

No consensus on stem cells

Problems with reproducibility bedevil research on adult stem cells, yet treatments using the cells are moving rapidly into human clinical trials. Those working in the field need to adopt more robust experimental approaches.

It was hailed as the dawn of a new era in the treatment of heart attacks. Three years ago, when researchers led by Piero Anversa of New York Medical College in Valhalla reported that heart damage in mice could be treated by injecting the damaged tissue with stem cells sucked from the animals' own bone marrow, hopes surged of repairing human hearts in the same way.

Several clinical trials are now under way, triggered by the results in Anversa's original paper (D. Orlic *et al. Nature* **410**, 701–705; 2001). But this week's issue of *Nature* contains two reports of failures to replicate the findings that should give those who are running these trials pause for thought (see pages 664 and 668). Unfortunately, this is not an isolated incident. Over the past few years, several teams have claimed that adult stem cells can develop into a wide variety of tissues — only for other researchers to fail to obtain the same results.

Being able to repair damaged tissues using a patient's own cells is an attractive therapeutic prospect, and has been promoted vigorously by lobbyists opposed to the use of human embryonic stem cells. As long as a patient's stem cells are not genetically modified, there are currently relatively few regulatory hurdles to be overcome before testing their therapeutic potential in the clinic, at least in the United States. As a result, promising findings from animal experiments can move into the clinic before contrary results come to light.

Why are studies using adult stem cells so problematic? Are there systematic, technical problems inherent in the studies? Or are discrepancies between the results obtained by different groups merely more noticeable than in other fields because of the high profile of stem-cell research?

It's probably a bit of both. And this means that those working in the field must put their technical house in order. Many of the problems

with reproducibility are thought to arise from the technique of immunofluorescent microscopy, in which specific antibodies bearing tags that fluoresce under laser light are used to track the migration and growth of injected stem cells. Unfortunately, this technique is prone to artefact. Antibodies may react with cells other than their intended targets, and microscopy is a notoriously subjective business. What's more, some cells fluoresce of their own accord under laser light.

Stem-cell researchers should adopt more robust techniques that incorporate marker genes that render stem cells visible using fluorescence and conventional microscopy. Ideally, the results should be confirmed using different marker genes. This will take time and work to achieve, but should be pursued as a high priority.

Other possible sources of variability in animal experiments include the techniques used to isolate, culture and characterize adult stem cells, and the precise type of tissue damage that the experimental treatment is designed to fix. Here, there are no clear answers, other than meticulous documentation of the methods used.

Researchers working on adult stem cells are under pressure both from severely ill patients and from companies looking for a profit. But a balance needs to be struck. Scientists embarking on studies likely to evoke clinical interest should work together, wherever possible, to resolve contradictions. Sadly, Anversa's team and the researchers who have failed to replicate his work have yet to exchange samples to try to understand why their results are at odds.

Perhaps it's time that institutional review boards and regulatory agencies, which in many cases have seemed happy to let therapies involving adult stem cells move towards the clinic with little intervention, started demanding a more robust consensus from animal studies before allowing human trials to go ahead. ■

Some sobering thoughts

We pay too much attention to the health benefits of alcohol and neglect the devastating effects of excessive consumption.

Compared with the tobacco industry, the companies that provide us with wine, beer and spirits have a glowing reputation. They have promoted campaigns to discourage drink driving and to educate the public in 'sensible' patterns of drinking.

Indeed, such is the drinks industry's status as a valued stakeholder that when the world's biggest research agency dealing with alcohol-related health problems, the US National Institute on Alcohol Abuse and Alcoholism (NIAAA), was looking for a new director in 2002, a representative of the San Francisco-based Wine Institute served on the search committee.

One popular refrain for the drinks industry is that, in moderation, alcohol can improve health. When the US Department of Agriculture revised its Dietary Guidelines for Americans in 1995, for example, the wine industry lobbied successfully for mention of the beneficial effects of moderate alcohol consumption on cardiovascular disease.

But you're unlikely to have heard industry representatives explaining that the beneficial effects of moderate drinking are limited to a relatively small proportion of the population. Many of the rest of

us, lulled into thinking that our 'social' consumption of alcohol is good for us, are literally drinking ourselves to death (see page 598).

The industry's preferred message has found subtle echoes in the research agenda. From the mid-1990s, language in the reports accompanying spending bills passed by the US Congress urged the National Institutes of Health to support research into the effects of moderate alcohol consumption. This coincided with the first moves by the NIAAA to fund substantial research in this area. In 1996, grants given to study the risks and benefits of moderate drinking totalled \$2.2 million. Although that was just 1.5% of the agency's budget, enthusiastic media coverage has since ensured that the findings from these projects have captured disproportionate attention.

Given the enormity of the world's drink problem, there is an urgent need to refocus public scrutiny on the harm caused by alcohol. If this harm is to be reduced, policy-makers will need to continue working in partnership with drinks manufacturers. But they should also put the alcohol industry's message of 'a little is good for you' in its proper context. ■





Extended family
Bioethics panel calls
for study into fertility
treatments
p590



Growth industry
California sows
the seeds for
drug-producing rice
p591



Prime mover
Spanish researchers
pin their hopes on
new government
p592



Catching up
Realigned hook sees
long-line fishing of
swordfish resume
p594

Bush administration dismisses allegations of scientific bias

Geoff Brumfiel, Washington

The White House has issued a detailed rebuttal to claims that it is distorting scientific evidence to further its political agenda.

The statement was issued on 2 April by John Marburger, President Bush's science adviser. It offers a point-by-point response to a February report by the Union of Concerned Scientists (UCS) that accused the White House of "misrepresenting and suppressing scientific knowledge" (see *Nature* 427, 663; 2004).

The UCS report, which was accompanied by a statement from some 60 leading researchers, generated considerable national media coverage, and Republicans and Democrats grilled Marburger about the charges it contains in congressional hearings last month.

Now the president's beleaguered science adviser has hit back. "I hope this response will correct errors, distortions and misunderstandings in the Union of Concerned Scientists' document," he writes in the introduction to the 20-page rebuttal.

The White House argues, for example, that a section on climate change in a June 2003 Environmental Protection Agency (EPA) report was not suppressed, as alleged, but was omitted in favour of a reference to the more extensive Climate Change Science Program released the following month. "The Administration chose, appropriately, to present information in a single ... far more complete format," the statement says. And the Department of Agriculture sought to prevent the publication of an analysis on the dangers of airborne bacteria in February 2002 because its



Fresh perspective: John Marburger has issued a point-by-point rebuttal of critics' accusations.

author, James Zahn, a microbiologist then at the Agriculture Research Service in Ames, Iowa, "did not have any scientific data or expertise in the scientific area in question".

Marburger has received some backing from the administration's supporters in Washington. "I don't think this administration has been politicizing science," says Robert Walker, a lobbyist and former Republican chair of the House Science Committee. "My concern is with the UCS report: it's a political diatribe."

But Sidney Drell, former deputy director of the Stanford Linear Accelerator Center in California, says he is "disappointed" by the White House document. "Marburger denies there's a problem," he says, "but there is one."

Kurt Gottfried, a particle physicist at Cornell University in Ithaca, New York, and chairman of the UCS, says that "on the most important issues, Marburger's response doesn't really stand up". The section on the EPA climate report, for example, fails to acknowledge an internal EPA memo that said it was changed for political reasons, he says.

Zahn, who is now at Iowa State University, says that the statement is "missing the point" when it says that he lacked expertise in airborne bacteria. "I never purported to be an authority on microbial transfer," he says. "But it was an important finding and it needed to be reported." Zahn also says that his findings have now been confirmed by a group in the Netherlands. ■

Law paves way for concealed guns on campus

Rex Dalton, San Diego

The state of Utah has passed legislation that will compel its universities to allow concealed handguns on campus.

The law, signed on 25 March by state governor Olene Walker and due to take effect on 3 May, instructs institutions including the University of Utah in Salt Lake City to

allow the guns. For more than two years, the university has battled to maintain its exemption from a 1996 law that allows permit holders to carry concealed weapons in the state (see *Nature* 417, 212; 2003).

The passage of the new law has caused consternation among faculty members at the university, which is currently seeking

a president to replace Bernard Machen, who left in January to become president of the University of Florida at Gainesville.

Andrew Gitlin, a sociologist who is president of the university's academic senate, says that the legislation will deter many candidates. "The first question they will be asked is: will you fight this or not?" he says. ■

Biotech company fights to keep tips on data access private

Jim Giles, London

An agricultural biotechnology company is asking a British court to prevent an environmental group releasing guidelines that would help the public probe pesticide trials.

Friends of the Earth, the London-based environmental group, wants to publish its knowledge on how to gain access to data. But Bayer CropScience of Monheim, Germany, says this would reveal commercially sensitive material to competitors, and is seeking a court injunction to prevent publication.

The guidelines draw on the environmental group's four years of legal struggles with Bayer over glufosinate ammonium, a pesticide used with a transgenic maize developed by the company. The pesticide was cleared for experimental use a decade ago, after the UK government's Pesticides Safety Directorate examined Bayer's data on human toxicity and environmental impact.

When Friends of the Earth first asked the directorate for those data in 2000, Bayer objected on the grounds of commercial sensitivity. The environmental group gained access last May — but could view the information only at the offices of Bayer's lawyers and could not copy or remove the documents.

Now the group wants to publish guidelines on how to access such data in Britain, including an array of procedural options beyond those used in the glufosinate ammonium case.

Last October, Bayer applied for an injunction to stop this, claiming that the information would encourage lawbreaking by providing access to information it wants to be confidential. "It is very easy to access information — people just have to approach us," says Bayer spokesman Julian Little. "But under the Friends of the Earth guidelines we will lose control of the data."

Friends of the Earth lawyer Phil Michaels says that the law will continue to protect commercially sensitive content, such as details of the active ingredient in the pesticide. "This is an absurd situation," he says. "The information is already available to the public. We just want to tell people how to access it."

Neither side will discuss what the guidelines actually say, until the court hears the case, which Michaels says could happen by June. ■



Baby boom: techniques such as egg harvesting, and concern about their effects, are both on the up.

Ethics council calls for probe into assisted reproduction

Erika Check, Washington

The health impacts of techniques used for assisted reproduction should be investigated in a comprehensive study supported by the federal government, according to a report by advisers to President George Bush.

The study would examine whether assisted reproduction is associated with higher rates of health ailments in children, such as genetic abnormalities and birth defects, as has been suggested by some smaller assessments (see *Nature* 422, 656–658; 2003).

The call comes amid growing concern in the United States about the rapidly expanding and largely unregulated assisted-reproduction industry. During the past 25 years, a million couples have borne children through such techniques. Pressure is building for Congress to regulate some aspects of the industry, such as the fact that new techniques do not require regulatory approval before being used in patients.

On 30 March, the President's Council on Bioethics, chaired by University of Chicago ethicist Leon Kass, issued a report on the topic — *Reproduction and Responsibility: The Regulation of New Biotechnologies*.

The report recommends that a more extensive investigation into the health effects of assisted reproduction should be incorporated into the National Children's Study (NCS), a major study of child health that has been in preparation since 2000.

Peter Scheidt, a paediatrician at the National Institute of Child Health and Human Development who directs the NCS, says that the study will almost certainly address assisted reproduction if it receives funding. But last month, an advisory committee to the child health agency warned that the NCS's sample population of 100,000

children would probably not be enough to tell whether assisted reproduction contributes to rare abnormalities.

The council's report also said that the government should study the health of egg donors and those who give birth to children through assisted reproduction. And it recommends that professional societies and the government should more tightly regulate the medical practice of assisted reproduction.

The bioethics council's work is always controversial, and this most recent report is no exception. At the council's meeting on 1 April, one of its members, endocrinologist Daniel Foster, questioned whether the removal of cell biologist Elizabeth Blackburn last month — and her replacement by non-scientists — would skew the council (see *Nature* 428, 4; 2004). "There is a concern that there is now going to be an imbalance," Foster says.

And although the report took no position on whether human embryos should be cloned to produce stem cells, some council members expressed concern that lawmakers might use the report to argue for a ban. In a personal statement appended to the report, council member Janet Rowley, a cell biologist at the University of Chicago, warned that she believed Congress might use some of the options outlined in the report "to do real damage to beneficial research and medical treatment".

Scientists' advocacy groups reacted cautiously to the report. "There is much in this report we can support," said Marian Damewood, president of the American Society for Reproductive Medicine, in a statement. "The call for follow-up on outcomes of children conceived with the use of reproductive technologies is very timely." ■

US biologist accused of robbing colleagues

Geoff Brumfiel, Washington

A biologist at the Dana-Farber Cancer Institute in Boston has been accused of swindling friends and colleagues out of more than half-a-million dollars to support a non-existent start-up company in China.

But the researcher, 38-year-old Weidong Xu, says that he raised the money honestly to support the study of severe acute respiratory syndrome (SARS) — only to lose it in a Nigerian investment scam.

Xu pleaded not guilty to stealing \$160,000 from six colleagues between July and September 2003, in a hearing at Roxbury District Court in Boston on 31 March. Prosecutors allege that he also took \$450,000 from almost 30 others.

Arnold Abelow, Xu's lawyer, says that his client fell prey to a Nigerian investment scam that promised to return \$50 million on an investment of \$600,000. He says that Xu always intended to fund a SARS-research company in his native China, but thought that the project could have a greater impact with the extra money. "He saw a fax in the Dana-Farber office from some people in Nigeria, and he fell for it," Abelow says.

In a statement, the Dana-Farber Institute said that its officials notified the police as



In the dock: Weidong Xu claims he lost funds for SARS research in a Nigerian investment scam.

soon as they were aware of the issue. When the police arrived to interview Xu, the court was told, they found him at a cafeteria arguing with another employee, who said he had given Xu more than \$5,000. Xu was immediately arrested; he is being held at the Suffolk County Jail in Boston, pending a court hearing on 23 April.

Until last month — when the institute fired him for reasons that it has not disclosed, but says are unrelated to the criminal charges — Xu had spent a low-key but respectable career in AIDS research. He had been at the Dana-Farber Cancer Institute since 1999, when virologist Ruth Ruprecht hired him to look for antibodies that might lead to HIV vaccines.

In 1993, after receiving a master's degree from Beijing Agricultural University, Xu went to Washington State University in Pullman to pursue a doctorate in biochemistry. On completing it in 1997, he joined the laboratory of Flossie Wong-Staal, a prominent AIDS researcher at the University of California, San Diego. "He was good at the bench, especially at molecular techniques," says Wong-Staal, now chief scientist at San Diego-based biotechnology company Immusol.

As a young biologist, Xu had a keen interest in business, says Thipparthi Reddy, a microbiologist who worked alongside Xu in San Diego and co-authored papers with him. "He was a patents-oriented kind of guy, and he always used to say: 'I want to work for a company,'" says Reddy, now at Wayne State University, Detroit. ■

BOSTON HERALD

California edges towards farming drug-producing rice

Rex Dalton, San Diego

Plans to step up cultivation of rice that is genetically modified to yield pharmaceuticals are sparking fierce opposition from some environmental scientists and farmers in California.

On 29 March, the California Rice Commission decided by 6 votes to 5 to advise the state's food and agriculture department to let Ventria Bioscience of Sacramento grow about 50 hectares of the rice at sites near San Diego. The department is expected to make a final decision later this month — approval would mean the rice could be planted immediately for harvest in the autumn.

Ventria wants to grow two types of rice, one modified to produce the human protein lactoferrin, which can be used to treat anaemia, and the other yielding lysozyme, an antimicrobial that is used to fight diarrhoea. In each case, it plans to grow about 60 tonnes of rice that would be processed to make 600 kg of the protein.

The sites have been chosen because of their remoteness from commercial rice farms in northern California. But some farmers claim that the rice, which isn't meant for human or animal consumption,



Field of doubts: farmers fear modified rice may undermine the market for their conventional crops.

could still cross-pollinate with food crops via migratory birds or through the water supply. Critics also say that perceptions of such a risk could damage California's rice exports to Japan and Europe, which were worth about \$500 million last year.

"We are opposed to the use of food crops as platforms to produce chemicals," says Bryce Lundberg of Lundberg Family Farms in Richvale, which grows about 1,400 hectares of rice, some of it by organic methods.

Prospects of growing crops in the United States with transgenes added to produce drugs have suffered several setbacks, including a possible case of contamination

of soya beans and maize in Iowa in 2002, which resulted in a \$250,000 fine for the Texas-based producer, ProdiGene.

Ventria's crop has been billed as the first of its type in California — although the company has already been growing the rice experimentally in the state for about five years. It also has experimental plots in Hawaii.

The company's chief executive, Scott Deeter, says that the firm grew 16 hectares of the rice last year at undisclosed locations in California. The US Department of Agriculture permitted these experiments, on condition the rice be grown at least 30 metres from conventional rice. ■

C. O'REAR/CORBIS

Web links leave abstracts going nowhere

John Whitfield, London

The Digital Libraries Initiative is a major US research project into electronic archiving. But when you click to its homepage from the website of one of its sponsors, the US National Library of Medicine, you get a message that's increasingly familiar to researchers: "URL not found".

Now a study by Jonathan Wren, a bioinformatician at the University of Oklahoma in Norman, has revealed the scale of the problem (J.D. Wren *Bioinformatics* 20, 668–672; 2004). He found that nearly one-fifth of the websites mentioned over the past decade in abstracts on Medline, the clearing-house of papers used by biomedical researchers, have disappeared.

"The web is becoming a more prevalent source of support for research," says Wren. "If that support is lost, it can make the work impossible to replicate."

Wren was prompted to study the problem after noticing a misspelt web address in the Medline abstract for one of his papers. In addition to the missing websites, Wren found that a fifth of the URLs in abstracts published between 1994 and April 2003 are available only intermittently. He also checked 33 FTP sites, of which only 12 still work. Wren hasn't quantified how important the dead links are, but he feels that they are probably significant as they were mentioned in the abstract.

And Robert Dellavalle, a dermatologist at the University of Colorado in Denver who is

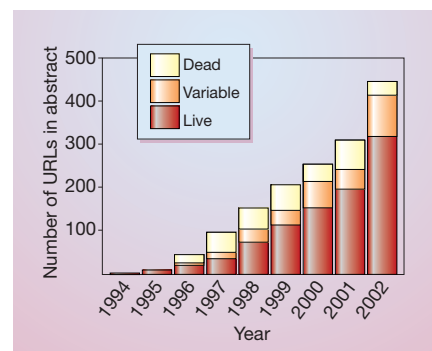
campaigning for better electronic archives, says that publishers' responses to the problem are inadequate. "Journals aren't doing anything to address the potential for electronic resources to disappear," he says. "It's amazing what doesn't exist — one of my own articles on digital preservation isn't there any more!"

Last year, Dellavalle's own survey found that about 12% of the Internet addresses cited in *The New England Journal of Medicine*, *The Journal of the American Medical Association* and *Science* were extinct two years after publication (R. P. Dellavalle *et al. Science* 302, 787–788; 2003).

Electronic resources in the physical sciences can be just as unstable, says Paul Ginsparg, a physicist at Cornell University in New York state, who runs the arXiv preprint service. "We've always insisted that arXiv submissions be complete and self-contained, and that external links be only to non-essential materials," Ginsparg says.

Dellavalle believes that journals should compel authors to archive the electronic resources they cite. As a minimum, he says, they should be required to print them out, and submit and keep copies. Additionally, he thinks, journals should require authors to submit online references to the Internet Archive (www.archive.org), a non-profit digital library project that has links to the world's largest library, the US Library of Congress.

One journal, *PLoS Biology*, is asking



Dead or alive? The availability of websites listed in Medline abstracts over the past decade.

authors to use the Internet Archive. Others are unsure how archives will be kept in the long term. "I don't think publishers have the resources to put everything somewhere it's going to last for ever," says Tony Delamothe, web editor of the *BMJ*. "If links die or get lost over time, that's just tough."

Maxine Clarke, publishing executive editor of *Nature*, says: "We do what we can to ensure that formal citations will stand the test of time, and we wouldn't let a significant part of a paper depend on a website."

Many commercial publishers, working through a collaboration called CrossRef, want to give electronic documents permanent code numbers that will prevent them from getting lost, even when URLs change. ■

Election promise gives hope to Spanish scientists

Laura Nelson, London

Researchers in Spain are optimistic that the newly elected government there will implement a bold pledge to double research spending. And they hope that it will also address problems that have fostered growing discontent on the nation's university campuses.

The socialist party of José Luis Rodríguez Zapatero, which scored a surprise election win on 14 March, promised in its manifesto to double Spain's €4-billion (US\$4.8-billion) annual research and development budget by 2008. It also said that it would recombine the science and technology ministry with the education ministry, and set up a research council to distribute grants on the basis of scientific merit.

Jaime Díez Lissavetzky, a chemist and socialist party member of parliament, who is helping to formulate the incoming government's science policy, says that the pledges will be implemented quickly.

In budget terms, science did not fare badly under the previous, conservative

government led by José Aznar. Its total research and development budget surged by an average of 13% each year from 1997 to 2003, according to European Commission figures — the second-fastest growth rate in the European Union.



Surprise victor: José Luis Rodríguez Zapatero (right) with outgoing prime minister José Aznar.

But scientists have been concerned about the low proportion of public money going into basic research and have protested over funding and the career prospects of young people.

The plan to merge the science and technology ministry back into the education department — the two were separated in 2000 — has drawn mixed responses from researchers, some of whom have been disappointed by what they regard as the ministry's focus on technology.

Emilio Gelpi, director of the Spanish research council's institute for biomedical research in Barcelona, says that the previous ministry was too involved in technological issues, such as telecommunications, to do much for science. Juan Belmonte, a geneticist at the Salk Institute in La Jolla, California, who collaborates closely with stem-cell researchers in Spain, says that he has mixed feelings about the combined ministry. But he adds that it has the potential to strengthen science in the universities, as well as its ties with industry. ■

Modellers deplore 'short-termism' on climate

Quirin Schiermeier, Lund

Projections of climate change in, say, Florida or the Alps carry more political weight than vague warnings about global warming. And for almost two decades, specialists in regional climate assessment have sought to make such projections.

But their success has been limited, a meeting of regional-climate modellers in Lund, Sweden, acknowledged last week. Our understanding of regional climate change will remain uncertain, the modellers said. And, some speakers suggested, policy-makers' expectations of precise local projections need to be dampened down.

"We're not yet at the promised level where regional climate models can really influence policy-making," Georgios Amanatidis, a scientific officer at the European Commission's research directorate, told the meeting. Participants admitted privately that the immediate benefits of regional climate modelling have been oversold in exercises such as the Clinton administration's US regional climate assessment, which sought to evaluate the impact of climate change on each part of the country.

Most of all, specialists in regional climate change feel that their work, while valuable in the long run, is misunderstood by those in search of immediate results.

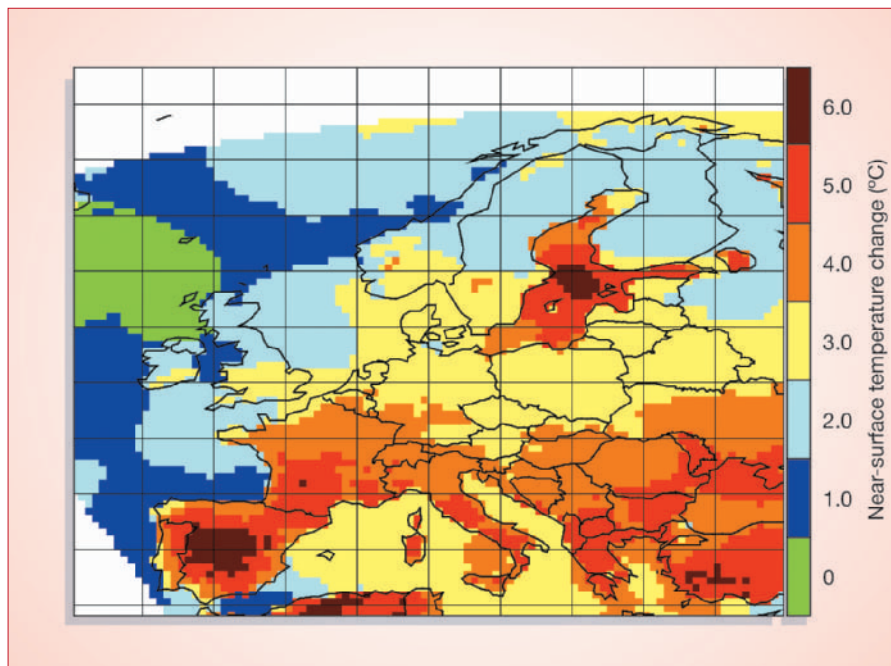
"Government and business would like us to predict what will happen in the next 20 years," says Colin Jones, a climate modeller at the Swedish Meteorological and Hydrological Institute in Norrköping. "Unfortunately, we can't — the natural variability of climate makes it next to impossible." The work is most useful for predicting trends beyond 2070, says Jones.

On target

As the planet heats up, scientists are trying to better understand the impact of global warming on continental, regional and local scales. This extra information could help planners concerned with monsoon regions, mountain areas or coastal zones to develop more targeted adaptation strategies against the adverse effects of climate change.

More powerful computers and better models are improving regional and local climate models, the meeting of 80 experts heard. Researchers are confident that they will generate information that will assist long-term planning in agriculture, flood protection and air-quality management.

The most ambitious regional-climate-change research effort is a European project called Prudence, which is receiving €3.5 million (US\$4.2 million) in European Union funding and will soon develop into a €17-million programme called Ensembles.



Scientists say that only long-range projections, such as this one for Europe in 2071–2100, count for much.

Prudence uses several models to assess European climate change. Under a pessimistic scenario for greenhouse-gas emissions, its results so far suggest that average temperatures in Europe will rise by 2–8 °C. Warming will be greater in central Europe and the Mediterranean than in the north, Prudence finds; European summers are likely to get drier, although rain heavy enough to cause flooding is likely to become more frequent.

But the size of these changes is uncertain — a message climate modellers say they find difficult to convey. Several participants in Lund said that their community struggles with communicating the stochastic nature of model-based predictions to grant-givers, decision-makers and the public. And many researchers would like to avoid the word 'prediction' altogether.

Regional accent

Climatologists first tried to track climate signals on a small spatial and temporal scale in 1987, at the US National Center for Atmospheric Research (NCAR) in Boulder, Colorado. They were trying to assess the vulnerability of Yucca Mountain, Nevada, where a nuclear-waste repository is planned, to changes in the water table.

Better computers have increased the use of regional models. "When I left the NCAR in 1997, it took one whole processor of the centre's supercomputer to run a regional simulation — now researchers are doing the same exercise on personal computers," says Filippo Giorgi, an atmospheric physicist now at the Abdus Salam International

Centre for Theoretical Physics in Trieste, Italy.

Giorgi, whom many regard as the founder of regional climate modelling, says he remains convinced of its value. But he does not try to hide the limitations of existing regional climate models.

All high-resolution simulations are based on global climate models. Thus, the weaknesses of the 'mother' models — around ten of which now run globally — rub off on regional simulations, which are based on the same equations representing the thermodynamics of the atmosphere. In regional models, these equations are solved simultaneously for a much larger number of grid points.

"If you don't believe in the value of global climate models then there's no point in downscaling them," says Giorgi. "But if you do — and global models do provide a quite consistent pattern of climate change — then it makes sense to translate global patterns into local information."

Thanks to Prudence, Europe is currently leading the field, Giorgi adds. "There is quite some envy over here of how well things are organized and funded in Europe," says Francis Zwiers, head of the Canadian Centre for Climate Modelling and Analysis in Victoria, British Columbia.

American and Canadian modellers are seeking funds for a similar collaboration of their own. "This would really come at the right time," says René Laprise, principal investigator at the Canadian regional climate modelling group in Montreal, "now that we are finally understanding what we're doing — some of the time!"

Transgenic crops find no takers in Britain or Australia

London Plans to commercialize genetically modified maize (corn) in Britain have been shelved just weeks after the government agreed to license the crop.

The maize, which has been engineered to resist a herbicide, was approved on 9 March following positive findings from a three-year experiment into the crop's environmental impact (see *Nature* 428, 107;

2004). But the company that developed it, Bayer CropScience of Monheim, Germany, said on 31 March that there were too many uncertainties to risk commercialization. The UK government has yet to clarify rules on the coexistence between transgenic and conventional crops, for example.

The agricultural-biotechnology industry also suffered two serious setbacks in the past two weeks in Australia. An application by Bayer and Monsanto, based in St Louis, Missouri, for a 3,000-hectare trial of transgenic oilseed rape in New South Wales, Australia, was rejected by the state

government last week, although three smaller studies were cleared. A week earlier, Western Australia announced a four-year moratorium on all genetically modified food crops.

Treasury lifts trade embargo on journal publishing

Washington The US government is lifting restrictions that stopped academic journals publishing research from certain countries under a trade embargo.

Last autumn, the Department of the Treasury informed the Institute of Electrical and Electronics Engineers (IEEE) that it needed a licence to publish work from Iran, North Korea, Sudan and Libya. The ruling drew fire from publishers, who claimed that the restrictions were unconstitutional, and left academic publishers split over whether or not to comply (see *Nature* 427, 663; 2004).

In a letter on 2 April to the IEEE, Richard Newcomb, director of the treasury's Office of Foreign Assets Control, announced the lifting of the restrictions based on a new understanding of the peer-review process. The letter effectively ends the debate about whether a licence is needed.

Arthur Winston, the IEEE's president, calls the decision "tremendous". He says that the institute will immediately return to its normal publishing process for all authors.

Fishermen line up to use turtle-friendly hook

San Diego After a three-year ban, long-line fishing of swordfish will resume off Hawaii next month — with the fate of endangered sea turtles hanging in the balance.

On 12 April, the US National Marine Fisheries Service published rules that will permit long-line fishing with a circular hook (pictured, right) that has been found to snag fewer turtles. The standard J-shaped hook kills up to 40,000 turtles each year.

Government monitors on every boat will check for turtle catches. If the boats catch more



than 16 leatherback or 17 loggerhead turtles in the season, long-line fishing will be halted.

Japan thwarts US bid to extradite biologist

Tokyo The Tokyo High Court has rejected a request from the United States to hand over a Japanese molecular biologist charged with economic espionage. The United States had sought to extradite Takashi Okamoto for stealing samples used in research on Alzheimer's disease from the Cleveland Clinic Foundation in 1999.

Reports last year suggested that Japan would extradite Okamoto, but on 29 March the court refused, deciding that Okamoto was not seeking profit for himself or another foreign entity. A charge of interstate transport of stolen materials was also dismissed.

The United States claimed that US\$2 million of funding went into producing the materials, but the court was not convinced they were worth more than the US\$5,000 needed to be in violation of the law.

French scientists' protests were 'justified', says Chirac

Paris French president Jacques Chirac performed an about-face last week, intervening to defuse protests by researchers over job and budget cuts. Speaking on television after the ruling conservatives' defeat in regional elections (see *Nature* 428,

454; 2004), Chirac disowned the research policies of his previous government, criticizing the "insufficiency of resources" and declaring the protests "justified".

In a government reshuffle on 31 March, Chirac nominated François Fillon as minister for education, higher education and research, and François d'Aubert as junior minister for research. Fillon and d'Aubert, who held the same positions during the 1990s, met research leaders last Friday, and



Research minister François Fillon will seek to ease French scientists' fears over job cuts.

an announcement on the protestors' key demand for posts for young scientists is expected this week. Chirac said he had told the government to re-examine this issue, and that the problem would be solved.

Creationist curriculum provokes Italian wrath

Rome School curriculum guidelines announced in Italy last month propose that children aged 11 to 14 do not need to be taught evolutionary theory in biology classes — but they should learn about creationism in religious studies, which are voluntary.

Leading Italian scientists have vowed to campaign against the guidelines, which were released by Italian education and research minister Letizia Moratti and are due to come into effect in October.

Carlo Redi, a cell biologist at the University of Pavia, complains that the guidelines adopt an overtly moralistic tone. For example, they stress health issues, such as the damage caused by illegal drugs and the benefits of a good diet, but neglect the basics of scientific theory and methods.

Redi and his allies are sending a letter of complaint to Moratti's ministry. The Accademia Nazionale dei Lincei, Italy's national scientific academy, will debate the issue on 22 April.

True colours

It's only when you begin to see the world as birds do — detecting light in the ultraviolet spectrum — that the full subtlety of their behaviour is revealed. Rex Dalton catches a glimpse.

With bug in beak, a parent starling squeezes into a dark hole in a tree eager to feed her gaping chicks. Turnaround time in the nest is crucial for the adult birds. They must bring as many squirming morsels as they can if their young are to fledge successfully. And in the struggle for survival, it pays to ensure that the fittest are fed first.

Swiss ecologists have now discovered how these birds distinguish their nestlings. The skin and the inside of the gaping beak of starling chicks (*Sturnus vulgaris*) reflect ultraviolet (UV) light, which the parents can see even in the dark nest. The researchers have even shown that nestlings with stronger immune systems reflect more UV — which would promote the survival of the fittest. “The nestling that stands out will be fed first,” says Philipp Heeb, a behavioural ecologist at the University of Lausanne, who led the team.

Most birds' eyes are thought to be sensitive to radiation with wavelengths between 320 and 400 nanometres, in the near-UV spectrum, to which we are blind¹. That's because birds have an extra class of photoreceptor cone cell in their retinas, in addition to the three — sensitive to red, green and blue light, respectively — that we possess.

Although this sensitivity was first discovered in the early 1970s (ref. 2), behavioural ecologists have only started to investigate the consequences of birds' UV vision in the past decade. And they are revealing just how much was missed by earlier studies. “Avian vision is much more complex than we ever realized,” says Rick Prum, an evolutionary ornithologist at Yale University in New Haven, Connecticut.

Birds use their UV vision for many things. Kestrels (*Falco tinnunculus*) use the tell-tale UV reflection of vole urine, left as scent marks in the environment, to home in on their prey³. Blue tits (*Parus caeruleus*) similarly seem to use their UV vision while foraging, to help detect camouflaged caterpillars⁴.



Feed me: Philipp Heeb (below) has shown that starling chicks reflect UV light as a signal to parents.

A sensitivity to UV may also have a role in egg recognition: the UV reflectance of the eggs of the African red-chested cuckoo (*Cuculus solitarius*) and those of the birds whose nests they parasitize turn out to be very closely matched⁵. So eggs that don't look especially similar to humans may look

sufficiently alike to a bird's eye to fool the cuckoo's victims.

But most of the behavioural work on bird UV vision has focused on the animals' own coloration, and its role in communication between members of the same species. Naturalists have long been fascinated by birds' bright colours, which are involved in camouflage, distinguishing genders and mate selection. But without UV vision, human birdwatchers have been missing an important part of the show. “Birds look great to us, but they look a lot better to themselves,” says Prum. “It's humbling.”

Rhapsody in blue

In 1996, visual ecologists Andrew Bennett and Innes Cuthill of the University of Bristol, UK, and their colleagues showed for the first time that a bird uses UV vision in choosing a mate⁶. Studying zebra finches (*Taeniopygia guttata*), the researchers put females in an arena in which they could hop towards various males. By deploying filters for different wavelengths among the birds — and in some experiments ornamenting some of the males with UV-reflecting leg-bands — the Bristol team showed that the females use UV reflectance to judge the relative attractiveness of different males.

Male and female zebra finches at least look different to the human eye. But in some other species, UV coloration is all that appears to separate the genders. Male and female blue tits look identical to us — but things look very different to them. In





Life in the ultraviolet: the zebra finch (above) gave the first clue to how birds use their UV vision in mate choice; female blue tits prefer males (left) with crests that look brightest in the UV spectrum.

separate studies, researchers led by Sarah Hunt, a member of Bennett and Cuthill's Bristol team, and Staffan Andersson of Gothenburg University in Sweden, have found that male blue tits differ from females in the UV reflectance of the blue feathers on the crest of their heads^{7,8}.

Indeed, the males' crests are brighter in the UV spectrum than they are at visible wavelengths. And Hunt and her colleagues went on to show in laboratory experiments that females prefer males with the brightest UV crests. Male blue tits, the Bristol researchers suggested in their paper, are actually "ultraviolet tits".

Such experiments depend on measuring the reflectance of the bird's plumage at various wavelengths using a hand-held spectrophotometer. Once spectrophotometers were bulky and expensive, and could only be used in the lab, but the latest equipment is the size of a cigarette pack and connects to any laptop computer. "It used to take all day to get one reflectance reading, but now it takes only a few minutes," says Bennett.

Skin deep

A bird's UV coloration goes deeper than its plumage. Behavioural scientists have long realized that chicks' gaping red and yellow mouths serve as signals that induce their parents to provide food. But a signal that appears red and yellow to us looks a different colour — one beyond our experience — to its intended recipients.

"A signal that appears red and yellow to us looks a different colour — one beyond our experience — to its intended recipients."

Last year, Hunt and her colleagues measured the UV reflectance of the gapes of chicks from eight different bird species, including the blue tit, house sparrow (*Passer domesticus*), barn swallow (*Hirundo rustica*), blackbird (*Turdus merula*) and reed warbler (*Acrocephalus scirpaceus*)⁹. In all cases, the nestlings' gapes were highly reflective in the UV spectrum. "I was surprised," says Bennett, "especially as it was consistent across all the species examined." What's more, chicks from species that nest in dark crevices showed especially bright UV reflectance from their 'flanges', the rim around the edge of the mouth.

Heeb's unpublished work, presented last October at the 2nd European Conference on Avian Colour Vision and Coloration at the Museum of Natural History in Paris, has shown for the first time that chicks use their body skin, as well as their gapes, to signal to their parents in the UV spectrum. Heeb's team first demonstrated that the reddish-brown skin of starling chicks reflected large amounts of UV light, and then went on to apply a UV-blocking jelly to some nestlings, and a control gel that didn't filter UV to others. The parents subsequently gave more food to the control chicks.

If chicks' UV skin reflectance is what evolutionary biologists call an 'honest' signal, then nestlings better equipped in the darwinian struggle for survival ought to reflect more UV light. To investigate, Heeb and his colleagues injected chicks with phytohaemagglutinin, a protein that provokes an immune reaction. By measuring the swelling around the site of the injection, the researchers showed that nestlings that reflected the most UV light had the strongest immune systems.

The chicks' UV reflectance may be due to the arrangement of fibres of the protein collagen in their skin. Prum and his colleagues have shown that this is the case for the blue facial skin patches worn by male asities, birds from Madagascar, during the breeding season¹⁰. In the next phase of their research, the teams at Bristol and at Yale intend to look further into how the structure of feathers and skin in various birds produces the UV reflections.

As researchers try to assemble the picture of how birds use their hues for survival and reproduction, some wonder whether we also need to rethink studies of other groups of animals. Mammals, for instance, do not see in the UV, but photoreceptor cells have evolved differently in various species to respond to different wavelengths of light. Even our close primate relatives may see the world in a very different light to us.

The lesson from the past decade of studies on bird behaviour is: if you want to get inside an animal's mind, it helps to see the world through its eyes.

Rex Dalton is Nature's US West Coast correspondent.

1. Bennett, A. T. D. & Cuthill, I. C. *Vision Res.* **34**, 1471–1478 (1994).
2. Huth, H. H. & Burkhardt, D. *Naturwissenschaften* **59**, 650 (1972).
3. Viitala, J., Korpimäki, E., Palokangas, P. & Koivula, M. *Nature* **373**, 425–427 (1995).
4. Church, S. C., Bennett, A. T. D., Cuthill, I. C. & Partridge, J. C. *Proc. R. Soc. Lond. B* **265**, 1509–1514 (1998).
5. Cherry, M. I. & Bennett, A. T. D. *Proc. R. Soc. Lond. B* **268**, 565–571 (2001).
6. Bennett, A. T. D., Cuthill, I. C., Partridge, J. C. & Maier, E. J. *Nature* **380**, 433–435 (1996).
7. Hunt, S., Bennett, A. T. D., Cuthill, I. C. & Griffiths, R. *Proc. R. Soc. Lond. B* **265**, 451–455 (1998).
8. Andersson, S., Örnborg, J. & Andersson, M. *Proc. R. Soc. Lond. B* **265**, 445–450 (1998).
9. Hunt, S., Kilner, R. M., Langmore, N. E. & Bennett, A. T. D. *Proc. R. Soc. Lond. B* **270**, (Suppl.) S25–S28 (2003).
10. Prum, R. O., Torres, R., Kovach, C., Williamson, S. & Goodman, S. M. *J. Exp. Biol.* **202**, 3507–3522 (1999).



The demon drink

A little alcohol can be good for you. But, collectively, our drinking habits are seriously damaging our health. Can public-health officials promote the message of moderation? Helen Pearson investigates.

Alcohol and tobacco are the terrible twins of public health. Both increase the risk of cancer and other life-threatening diseases. Both are promoted aggressively by a powerful industry. And both can be horribly addictive.

Yet while packets of cigarettes bear stark health warnings and smokers have to skulk outside their workplaces to indulge the habit, alcohol has largely escaped such vilification. Indeed, as public-health officials concentrate on our expanding waistlines, you could be forgiven for thinking that drinking doesn't pose serious health problems. "It's even getting a softer ride than food these days," says Robin Room, director of the Centre for Social Research on Drugs at Stockholm University in Sweden.

But the latest figures on the global burden of disease suggest that it's time to rethink. According to a new report from the World Health Organization (WHO), the harm caused by alcohol nearly equals that from smoking¹. Add in the social consequences of

drinking, including wrecked families and lost days at work, which were excluded from the WHO report, and alcohol should arguably be the greater concern.

There are some small signs that alcohol is surfacing on the global health agenda. For the first time in 20 years, national delegations at May's World Health Assembly, the annual meeting in Geneva, Switzerland, that sets the WHO's agenda, will consider a resolution that proposes making alcohol a priority for action. "It is a sign of something good," says Jürgen Rehm, who studies public-health policy at the University of Toronto in Canada.

Mixed messages

One reason that alcohol has not been targeted so effectively by public-health campaigns is that its detrimental effects are less clear-cut than those of smoking. Tobacco is 100% bad: even one cigarette does you harm. But with alcohol, the picture is much more complicated. Although a handful of

researchers still question the idea that moderate drinking is beneficial, the consensus from epidemiological studies is that alcohol can improve health if taken in small doses — primarily by reducing the risk of heart attacks in older people (see 'A little of what you fancy', overleaf).

On the surface, this health message seems simple, but its consequences are not. Most of us like a drink, many valuing alcohol as a social lubricant and reliever of stress. The trouble, however, is that far too many of us like to take much more than one drink. Bombarded by stories in the media portraying the medicinal effects of moderate consumption — but with no consistent message on what defines moderation — many drinkers are unclear about the damage they're doing to their health.

Against this background, health authorities struggle to explain that the health benefits of alcohol are largely limited to a few individuals. Indeed, if you look at the issue at the population level, it's clear that we need to

Drinking ourselves to death



cut back.

"It does present a conflict in terms of alcohol policy," says Thomas Babor, a health-policy researcher at the University of Connecticut School of Medicine in Farmington.

Epidemiological evidence that alcohol could actually improve health first made an impact in the 1970s. Since then, studies from all over the world have shown that moderate alcohol intake — one or two drinks a day — lowers the risk of coronary heart disease by 20–30% for older men and women².

Little payback

This basic message is now well known. But the cardiovascular gains from alcohol must be weighed against its numerous evils. Even small amounts of alcohol increase the risk of injury and boost the chances of developing about 60 diseases, including several cancers, liver cirrhosis and neuropsychological disorders. When these are fed into the epidemiological equation, only men over the age of 45 and women over 55 seem to lower their overall health risks by moderate drinking³.

Translated into guidelines for sensible drinking, this means that a middle-aged teetotal man, with high cholesterol or a family history of heart disease, would probably benefit by including a single drink a day with dinner. He might, however, achieve the same or better payback by starting to exercise or adopting a healthier diet. And for older people already following an exemplary lifestyle, the benefit reaped from a daily drink is probably marginal.

Those in their 20s and 30s, meanwhile, may start accruing heart benefits from a daily drink, but at this age the threat of accidental injury equals the negligible risk from heart disease, so even moderate drinking seems to offer little payback.

Experts say that an individual's ideal level of consumption should be tailored to

their age, medical history and physique, and is preferably prescribed in a doctor's office. This makes it difficult to convey public-health advice to the population as a whole. "I don't think any indiscriminate advice is suitable — it can't be a 20-second sound bite," says cardiologist Arthur Klatsky of the healthcare organization Kaiser Permanente in Oakland, California.

"The message that moderate drinking can promote health may perversely encourage some people into dangerous levels of consumption."

Another difficulty in conveying advice on healthy alcohol consumption is the lack of a standard definition for 'a drink'. The alcohol content of any drink varies with its volume and alcohol concentration, so that guidelines advocating a 'moderate' intake are often inconsistent within and between countries. This was highlighted in a December 2003 report⁴ from the International Center for Alcohol Policies in Washington DC. This revealed that a 'standard' drink in Britain contains 8 grams of ethanol, in the United States it contains 14 g and in Japan 19.75 g.

Even if national health guidelines spell out the limited benefits of moderate drinking, the public tends to hear a rosier message. Experts pin part of the blame on the media, which, they say, tend to make a splash with studies showing alcohol's benefits, rather than explaining the relative risks. After a 1991 edition of the US TV programme *60 Minutes* on the health benefits of wine, trade magazines reported a hike in the next month's wine sales of more than 40%. In part, the soft ride given to alcohol may be linked to the fact that the media and other influ-

ential sectors of society tend to be drinkers. "The press lives in a very wet world — and so do the politicians," says Room.

The alcohol industry has also tried hard to disseminate the virtues of its products. In the mid-1990s, the US wine industry proposed bottle labels advertising the health benefits of moderate drinking. Some labels appeared but were quickly abandoned when the US Bureau of Alcohol, Tobacco, Firearms and Explosives provisionally banned such statements in 1999.

Industry representatives argue that they promote responsible drinking by broadcasting or contributing to public-health campaigns. At the same time, they acknowledge that the medical evidence on alcohol's health benefits has helped them persuade policy-makers of their product's appeal. "It distinguished us from the tobacco companies because we're not seen as inherently bad," says Stephen Whitehead, corporate affairs director with Allied Domecq, a spirits and wine producer based in London.

Global problems

Given the conflicting messages coming from health authorities, the media and the alcohol industry, it should come as no surprise that drinkers are confused. Clinical psychologist Nancy Piotrowski of Capella University in San Francisco and the Alcohol Research Group in Berkeley, California, has found that more than three-quarters of US healthcare providers did not define moderation in accordance with national guidelines. Her unpublished survey also sampled 130 drinkers in northern California, whose view of moderation was often influenced by their own alcohol consumption. "Many people define 'moderation' as what they do," Piotrowski says.

This means that the message that moderate drinking can promote health may perversely encourage some people into dangerous levels of consumption.

"Plenty of people use this message as an excuse to drink more alcohol," argues Ira Goldberg, professor of preventive medicine at Columbia University in New York.

At the population level, the hazards of alcohol are plain to see. Internationally, the highest disease load attributable to alcohol is found in the heavy-drinking former socialist countries of Eastern Europe and in Latin America (see Map, above). Rich developed



To reduce crime, some towns have banned street drinking.

countries have the next biggest problem, with public-health experts being especially worried about binge drinking among the young. Trends differ from country to country — consumption is rising in Britain, for instance, but falling in France and Italy.

In most developing countries, alcohol consumption is still relatively low, but it is climbing steeply, particularly in Asia, driven by economic growth and aggressive marketing. Without intervention, experts predict a future wave of alcohol-related health problems across the developing world.

Governments are mainly trying to tackle the threat posed by alcohol through school education or public-health messages in the media. But given the difficulty of imparting the message of moderation, most experts believe this approach is unlikely to have much effect. “We don’t think it’ll be successful on its own,” says Michael Marmot of University College London. He chaired a working group of Britain’s Academy of Medical Sciences that last month⁵ reported on the need to curb heavy drinking in Britain — where overall alcohol consumption has risen by 50% since 1970.

Strong measures

What does work, Marmot’s working group concluded, after reviewing the evidence from studies conducted worldwide, are more forceful measures to curb consumption. These include raising the minimum legal drinking age, cutting the number of liquor shops, and restricting advertising and the hours during which drink can be sold. Perhaps unsurprisingly, the price of alcohol has a very powerful influence on total consumption (see Graph, right). Most people, it seems, will curtail their drinking if hit hard enough in the pocket.

Experts point to the last few years of the Soviet Union, when excessive drinking was devastating the nation’s health. In 1985, Mikhail Gorbachev’s government clamped down on the production of illegal alcohol to curtail drunkenness. The move slashed consumption from 14.2 to 10.5 litres of pure alcohol per person per year — and is estimated to have cut the number of alcohol-related deaths from accidents, poisoning and violence by 11% over five years⁶. But in present-day Russia, alcohol consumption and related mortality have climbed back to pre-1985 levels.

In a one-party state the Soviet leaders didn’t have to worry too much about a public backlash. But few democratic politicians are keen to impose higher taxes on alcohol and restrict its sale — there are few votes to be gained, and many potentially to be lost, from such actions.

Marmot’s working group acknowledged this difficulty, but pointed to the success of anti-drink-driving campaigns as evidence that initially unpopular moves can be successful. The key to success, the academy’s

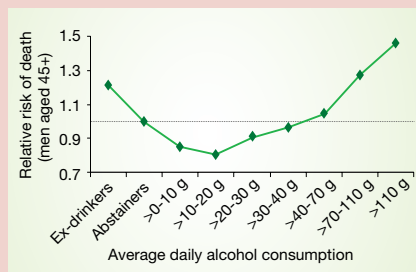
A little of what you fancy

Today, the idea that moderate drinking can be beneficial to health is well established. But in the 1980s and 1990s, some sceptics suggested that this effect might be a statistical mirage⁷. Teetotalers, they argued, include ex-alcoholics, the elderly and ill. So the apparent health benefits of moderate drinking could in fact be an artefact.

Some epidemiologists still wonder whether abstainers’ relatively poor health is linked to some factor not directly related to their rejection of alcohol. “Non-drinkers are a different type of person,” says Ian White, who has carried out epidemiological studies of drinking at the UK Medical Research Council’s Biostatistics Unit in Cambridge.

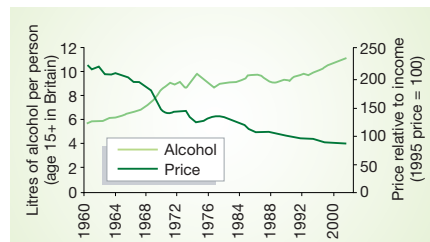
But more recent epidemiological analyses have adjusted, so far as is possible, for confounding effects, and still find health benefits from moderate drinking^{8–10}. They are backed by studies showing that alcohol benefits the body by boosting levels of ‘good’ cholesterol and blood proteins that are signs of a healthy heart¹¹. “There’s not much that can be done to make it more convincing,” says epidemiologist Eric Rimm of the Harvard School of Public Health in Boston.

Another issue that has mostly been laid to rest is that wine is better than beer or spirits —



a phenomenon labelled the ‘French paradox’ because it was thought to explain why citizens of this nation tend to escape heart disease despite an artery-clogging diet. Among many studies contradicting this idea, one of Czech beer drinkers found that those who imbibed four to nine litres each week had the lowest risk of heart disease, compared with those who drank more or less¹².

The remaining debate is over what constitutes a healthy pattern of drinking. Seven drinks a week can be supped one per day with dinner or downed together in a Friday night. The weight of studies suggests that binge drinking in this way does not protect from heart disease in the same way as spreading it out over several days¹³. “But there’s a little uncertainty there,” says Rimm.



report argued, will be to make the public more aware that measures to restrict the sale of alcohol and increase its price can relieve the unwanted consequences of excessive drinking, including violent crime. “We’d like the public to demand these things,” says Marmot.

But in Britain, the academy’s message seems to be falling on deaf ears. Just days after the report came out, the UK government released its long-awaited strategy on alcohol. It stressed partnerships between government and the drinks industry to promote sensible drinking, but proposed no increase in the cost of alcohol. At the same time, plans are afoot to extend the time over which British bars are allowed to stay open.

If harder-hitting measures are to be introduced, some experts believe lessons can be learnt from the campaign against tobacco. In that case, the battle was taken to the WHO, which last year adopted the Framework Convention for Tobacco Control. Among other measures, this helps countries to increase

taxes on tobacco without contravening international trade agreements.

But the convention was only adopted after a long and bruising battle with the tobacco industry, and public-health experts wonder whether the WHO will have the stomach for a similar protracted fight with the drinks industry. Some fear that alcohol-related health problems will have to reach crisis point before governments and international agencies take concerted action. “Sometimes problems have to escalate before they’re convinced to do something drastic,” says Babor.

Helen Pearson works in New York for Nature’s online news team.

1. Babor, T., Rehm, J. & Room, R. *Alcohol’s Role in the Global Burden of Disease: Implications for Alcohol Policy* (World Health Organization, Geneva, in the press.)
2. Marmot, M. G. *Int. J. Epidemiol.* **30**, 724–729 (2001).
3. White, I. R., Altmann, D. R. & Nanchahal, K. *Br. Med. J.* **325**, 191–197 (2002).
4. *International Drinking Guidelines* (International Center for Alcohol Policies, Washington DC, 2003).
5. *Calling Time: The Nation’s Drinking as a Major Health Issue* (Academy of Medical Sciences, London, 2004).
6. Nemtsov, A. V. *Addiction* **97**, 1413–1425 (2002).
7. Shaper, A. G. *Br. J. Addict.* **85**, 837–847 (1990).
8. Corrao, G., Rubbiati, L., Bagnardi, V., Zamboni, A. & Poikolainen, K. *Addiction* **95**, 1505–1523 (2000).
9. Rehm, J., Guttjahr, E. & Gmel, G. *Contemp. Drug Problems* **28**, 337–361 (2001).
10. Mukamal, K. J. *N. Engl. J. Med.* **348**, 109–118 (2003).
11. Rimm, E. B., Williams, P., Fosher, K., Criqui, M. & Stamper, M. J. *Br. J. Med.* **319**, 1523–1528 (1999).
12. Bobak, M., Skodova, Z. & Marmot, M. *Br. J. Med.* **320**, 1378–1379 (2000).
13. Puddey, I. B., Rakic, V., Dimmitt, S. B. & Beilin, L. J. *Addiction* **94**, 649–663 (1999).

Scientific aid to Brazil is strangled by red tape

The cost of importing donated equipment can be more than its original purchase price.

Sir—Young Brazilian scientists not only have to pay high prices for equipment, as reported in your News story “High prices of supplies drain cash from poorer nations’ labs” (*Nature* **428**, 453; 2004), they receive little funding from government agencies, and encounter obstacles if they obtain international support. For instance, when a Brazilian institution has valuable equipment donated by laboratories in the developed world, getting the equipment through customs is a surreal experience. Not only does it require tremendous amounts of paperwork; in some cases release from customs can take more than a year, during which time storage is charged. Consequently, the cost of importing scientific equipment to

Brazil is often higher than the cost of the equipment itself in the developed world.

What is at stake here is more than specific items of equipment. This situation risks undermining the creation in Brazil of new research groups led by young scientists, trained abroad in the most up-to-date techniques. In the United States, for example, the Pew Latin American Fellows Program awards junior biomedical scientists US\$35,000 at the end of their US postdoctoral training, to help establish laboratories back home. The benefits of such schemes are many, but they will falter without a different policy towards foreign scientific trade and donations.

The nomination of a new minister of

science and technology, Eduardo Campos, offers some hope. Although state and municipal institutions are also responsible for delays in importing goods, the possibility of reform is mostly in federal hands. President Luis Inacio Lula’s administration, which was elected with a mandate for change, should give immediate attention to these matters.

Stevens Kastrup Rehen

Department of Molecular Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, ICND118, La Jolla, California 92037, USA

Signed on behalf of:

A. Muotri, A.-M. Landeira-Fernandez, J.-G. Abreu, M. da Silva Almeida, M. Guimarães, R. Mohana-Borges, S. Ribeiro.

Full addresses are available from the corresponding author.

Ugly truths should not stop Pakistan’s reforms

Sir—Your Editorial about science reform and alleged sales of nuclear technology to Iran, “Good and bad in Pakistan” (*Nature* **427**, 379; 2004), mentions both good and bad, but stops short of citing the ‘ugly’. Yet there are two ugly truths involved.

The first is the state of international politics, which keeps shifting its goalposts, and accordingly either ignores the obvious or sometimes searches for the obscure. Witness, for example, the newly discovered activities of Abdul Qadeer Khan, Pakistan’s former chief nuclear scientist. It beggars belief that a solo scientist would be capable of exchanging nuclear know-how for missiles without the active participation of the army, and hence the government. It is getting clearer now that key players in the international community, including the United States, knew a lot about these activities but chose to remain quiet (see, for example, the *International Herald Tribune* of 21 April 2003; www.iht.com/articles/93839.html).

The second ugly truth is the fact that Pakistan never signed the nuclear non-proliferation treaty, and therefore the International Atomic Energy Agency can do very little about Pakistan’s proliferation activities in any case.

Your Editorial was perhaps too quick in predicting repercussions from these events on science in Pakistan, which has been making progress recently.

It is more probable than not that President Musharraf, in order to gain the confidence of the country’s wider scientific community and to prove that the matter was nothing but an isolated incident,

would continue to finance reforms.

Debasish Debnath

Department of Surgery, Medical School, University of Aberdeen, Foresterhill, Aberdeen AB25 2ZD, UK

Gulf Stream safe if wind blows and Earth turns

Sir—Your News story “Gulf Stream probed for early warnings of system failure” (*Nature* **427**, 769; 2004) discusses what the climate in the south of England would be like “without the Gulf Stream”. Sadly, this phrase has been seen far too often, usually in newspapers concerned with the unlikely possibility of a new ice age in Britain triggered by the loss of the Gulf Stream.

European readers should be reassured that the Gulf Stream’s existence is a consequence of the large-scale wind system over the North Atlantic Ocean, and of the nature of fluid motion on a rotating planet. The only way to produce an ocean circulation without a Gulf Stream is either to turn off the wind system, or to stop the Earth’s rotation, or both.

Real questions exist about conceivable changes in the ocean circulation and its climate consequences. However, such discussions are not helped by hyperbole and alarmism. The occurrence of a climate state without the Gulf Stream any time soon — within tens of millions of years — has a probability of little more than zero.

Carl Wunsch

Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA

Poet described stars in Milky Way before Galileo

Sir—It is commonly accepted that Galileo in 1609 was the first to appreciate the discrete nature of the stars in the Milky Way, by training his telescope on that part of the heavens. Your review of Francesco Bertola’s book *Via Lactea* (*Nature* **427**, 489; 2003) notes Bertola’s suggestion that the true nature of the Milky Way may have been known before Galileo’s observation.

There is evidence to support this view. Sonnet 31 of Thomas Watson’s *Hekatompathia* (1582) describes the Milky Way as being composed of a huge number of discrete stars: “That can not tell how many starres appeare/ In part of heav’n, which Galaxia hight” — “Galaxia” is identified as the Milky Way in the notes to the poem. Sonnet 31 also appears in an earlier version of the *Hekatompathia* called *Looking glase for Looers*. Watson, it seems, knew of the discrete nature of the stars before 1582.

Sebastian Verro also describes the Milky Way as a collection of discrete stars in his 1581 *Physicorum Libri X*, Book II, chapter 17, page 31: “We now refer to the glorious Galaxia, which is also called the Milky Way. It is a chaos of minute, brightly shining stars, as if a fog or mist, which traverses the sky in an oblique path” (our translation).

Perhaps Watson and Verro gained their knowledge of the Milky Way through the use of earlier instruments than Galileo’s telescope, such as the perspective glasses of the sixteenth-century English natural scientists Leonard and Thomas Digges.

Eric Lewin Altschuler*, **William Jansen†**

**Mount Sinai School of Medicine, 1425 Madison Avenue, Box 1240, New York, New York 10029, USA
†2611 18th Ave (G), Forest Grove, Oregon 97116, USA*

Clarifying consciousness

Work is under way to relate the mind to observable neurobiological events.

The Quest for Consciousness: A Neurobiological Approach

by Christof Koch

Roberts: 2004. 448 pp. \$45, £29.99

Wider than the Sky: The Phenomenal Gift of Consciousness

by Gerald M. Edelman

Yale University Press: 2004. 185 pp. \$24

Jean-Pierre Changeux

The relationship between body and soul has been fiercely debated since Ancient Greece, first by philosophers and then by scientists. But until the 1980s it was not considered 'scientifically correct' for physiologists to mention the word 'consciousness' in a refereed paper. Christof Koch's book *The Quest for Consciousness* shows that this period is definitively over. The many facets of consciousness that philosophers have put forward are still not yet understood, but Koch's book represents an exceptionally rich presentation of existing experimental work, accompanied by many cogent, if still informal and debatable, theoretical statements.

From the start, the reader is taken on an in-depth exploration of the most recent developments in the biology of consciousness, guided by Koch and Francis Crick, with whom Koch has collaborated for the past 20 years. The outcome is, in my view, exceptional.

This book and a few others, in particular Gerald Edelman's *Wider than the Sky*, also reviewed here, bring us to the stage where, in my opinion, objective physiological recordings and measurable physical parameters can be causally linked with subjective experience. This is a strong statement. Koch is more cautious, stating that his quest is for the neuronal correlates of consciousness. His aim is to establish an "explicit correspondence" between mental and neuronal events. But, in my opinion, causal determination, despite all its difficulties, is required. Koch's position therefore seems less bold than that set out in Crick's own book *The Astonishing Hypothesis* (Simon & Schuster, 1994): "You are in fact no more than the behaviour of a vast assembly of nerve cells and their associated molecules."

The explosive development of new methods of brain exploration, in particular brain imaging, as well as imaginative cognitive experimentation, has created an important experimental literature on consciousness, which Koch's book covers in a vivid and fluent style. He has deliberately not covered some important issues, such as the self-referential aspect of consciousness and the role of language and emotions. Instead, the focus is on perception, particularly visual



The 'exotic landscape' of our perceptions, reflected in this painting by Henri Rousseau.

perception. Koch traces the diverse routes from the retina through the visual cortices to the frontal cortex and then to the motor areas, defining step by step the necessary conditions for the neuronal correlates of consciousness. This bias towards an input-output view of information processing neglects the role of spontaneous states of neural activity and contrasts with the top-down projective style of the brain, which constantly anticipates the evolution of the outside world — this should, in my opinion, be an essential part of any theory of consciousness.

Perhaps Koch's most controversial proposal is that "zombies could be living among us". According to Crick and Koch, these fictitious creatures would be "devoid of any conscious experience" and yet would have behaviour identical to that of their normal, conscious counterparts. This can only be a metaphor, not a serious model, yet it might help to illustrate the notion that specialized sensorimotor reflex processes (referred to as 'zombie agents') carry out routine missions in the absence of any direct conscious sensation or control.

This concept is reminiscent of Pierre Janet's "subconscious actions", as presented in his 1889 book *L'automatisme psychologique*, which are "entirely absent from the vision of the subject, thus forcing ourselves to assume their existence outside consciousness". Janet, like Koch, based his argument on a clinical framework, including catalepsy, somnambulism, hallucinations and anaes-

thesia. It is fascinating to follow Koch's presentation of the latest clinical observations, together with abundant psychophysical experiments, in which reasonable attempts have recently been made to uncover neural signatures for conscious and non-conscious processing. Examples include the views from functional magnetic resonance imaging (fMRI) of the masking experiments of Stanislas Dehaene and colleagues, and Nikos Logothetis' dynamic electrophysiological recordings and fMRI images from binocular-rivalry experiments.

Koch also discusses the idea that the perception of motion consists of discrete processing epochs, or snapshots. He re-examines the classical paradox of the continuity of cinematographic vision, recently refreshed by Oliver Sacks' clinical observations of patients who see movement as a succession of still images during migraine episodes. This rather simple experimental paradigm deserves to be re-explored using the tools available today. Understanding this, as well as many of the other issues raised by *The Quest for Consciousness*, should tell us more about the access to, and persistence of, visual stimuli in consciousness.

Koch postulates that if perceptual representations are to achieve conscious status, they must be intimately linked to an intentional planning stage. He therefore introduces a requirement for projection to the front of the brain. Nevertheless, Crick and Koch follow Ray Jackendoff in postulating

that many aspects of high-level cognitive functions, such as decision-making, planning and creativity, are themselves hidden from consciousness: "You don't know your innermost thoughts." These thoughts are carried by "a little person, a homunculus, inside my head who perceives the world through the senses, who thinks, who plans and carries out voluntary actions" — another metaphor. This position implies that "you are aware only of the sensory representations associated with these mental activities."

This is reminiscent of a debate in the nineteenth century about the role of images in thought. In 1870, Hippolyte Taine compared the mind to "a polype of images", and recent ingenious experiments by Gordon Shepard and Stephen Kosslyn have confirmed the role of mental imagery as central to the substance of conscious thought. Yet it seems highly debatable that all the contents of consciousness are sensory. What about mathematical concepts and their creation? What about consciousness of our errors? It seems to me that many abstract non-sensory representations can also be conscious. But Koch's speculation does bring forward an important question: can we define the particular kind of conscious and non-conscious representations to which the prefrontal cortex contributes? I am convinced that these issues can now be investigated in neurobiological terms.

An important quality of *The Quest for Consciousness* is the book's attempt to compare the views of Koch and Crick to the work of others — in particular to Edelman's sophisticated framework, which is elegantly summarized in *Wider than the Sky*. Koch's competition between coalitions of neurons fits well with Edelman's clusters of neurons and group selection by a massive feedback signalling loop. However, Edelman's position contrasts in another respect with that of Crick and Koch (and Dehaene and myself, for that matter), who posit well-defined neural architectures for consciousness. These structures appeared in the course of evolution — possibly from what Derek Denton refers to as "primal emotions", such as thirst, hunger and sex — and relate in particular to the expansion of the prefrontal cortex and its inhibitory power. A further distinction between the view of Crick and Koch and that shared by Dehaene and myself is our emphasis on neural mechanisms of evaluation for actualized actions, but also self-evaluation for tacit plans. This crucial mobilization of reward systems in consciousness is missing from Koch's book.

In a field that is plagued by more philosophical than scientifically sound controversies, Koch's book is, on the whole, remarkably balanced. But it does not touch on one crucial issue: can a model of consciousness be formalized in mathematical terms? Is it realistic to conceive a computer model or a robot that implements the conscious versus

non-conscious processing of sensory stimuli? My own answer would be a resounding 'yes'. I hope that the next generation of neuroscientists, inspired by reading *The Quest for Consciousness*, will soon start planning their experiments. ■

Jean-Pierre Changeux is at the Molecular Neurobiology Laboratory, Institut Pasteur, 28 rue du Docteur Roux, 75724 Paris, France.

Keep it simple

Deep Simplicity: Chaos, Complexity and the Emergence of Life

by John Gribbin

Allen Lane: 2004. 272 pp. £18.99

Mark Buchanan

The search for simplicity is perhaps the most basic theme of all science. As the late social and political scientist Herbert Simon put it, the purpose of science is "to find meaningful simplicity in the midst of disorderly complexity". In his new book *Deep Simplicity*, John Gribbin explores this theme in the context of two great movements of modern science — chaos and complexity — and argues that the discovery of simplicity hiding behind surface complexities will soon explain the origin of life itself.

Gribbin suggests quite plausibly that humans — and, by implication, our societies — are among the most complex things in the Universe. At the atomic level, individual particles follow relatively simple physical laws. It is out of the interactions of many particles, and then of objects made of them, that complexity arises, producing conductors and liquid crystals, biomolecules, living organisms, ecosystems and human culture. On a larger scale, the world again becomes relatively simple, for in the interior of a large planet, or a star, "gravity crushes any structure out of existence". Complexity occupies a middle world, which is also, probably for good reason, our world.

The aim of the book is to explore how simplicity arises on this level, and how it can be identified. But first Gribbin establishes why complexity, or at least the appearance of complexity, should be expected.

The classical newtonian view of the fully predictable Universe dominated science for two centuries. But as Gribbin points out, this world view actually rested on a vast leap of faith — on the supposition that if scientists were clever enough to solve Newton's equations for any system of interacting particles, their solutions would be just as regular as the periodic motion of two bodies, reflected in the elliptic orbits of the planets about the Sun. In 1890, the French mathematician Henri Poincaré proved otherwise: that the resulting motion can be irregular and unpredictable, even when only three bodies are involved. "It

may happen," Poincaré wrote, "that small differences in the initial conditions produce very great ones in the final phenomena." This, in modern parlance, is chaos, and it implies — in the more general context of dynamical-systems theory — that scientific prediction over long periods of time is generally impossible.

Gribbin tells the story of the modern rediscovery in the 1960s and 1970s of Poincaré's insight. This is an exciting tale but has been told before, most notably in James Gleick's bestseller *Chaos* (Heinemann, 1988). On the positive side, the discovery of chaos reveals that many highly erratic phenomena, ranging from chemical reactions to fluctuations in biological populations, may actually arise from very simple underlying dynamics. This is one way that simplicity often lies behind complexity.

Gribbin then weaves the story of chaos together with more recent developments, and with a host of topics now gathered together under the term 'complexity science'. The book moves rapidly from spontaneous pattern formation to the mathematics of fractals and the idea of self-organized criticality, examining its relation to earthquakes, mass extinctions and a vast range of other prominently unpredictable phenomena. The book celebrates the contemporary emphasis, especially in physics, on seeking the explanation of complex phenomena through simple dynamical models of growth and evolution. The lesson is the same everywhere: what appears as surface complexity often has its origins in dynamical simplicity. Importantly, Gribbin points out that modern computers have played a central role in making the complexity sciences possible, altering not only the content of science but the way it is done.

Much of *Deep Simplicity* will be familiar to anyone who has read about chaos and complexity before, but Gribbin does his usually excellent job of making complicated ideas accessible to a broad readership, and the book would certainly make an excellent non-technical introduction to this way of thinking. One minor shortcoming is that the book could have been written in, say, 1998, and still contained virtually all the same material. This is a little disappointing, as the past five years have witnessed a flowering of the complexity sciences and their successful application to a broad range of scientific topics.

Gribbin is something of a phenomenon of science writing, judging from his prolific output over the past two decades. In *Deep Simplicity*, perhaps, he doesn't quite succeed in showing how chaos and complexity will soon "explain the origin of life itself". But he breathes life into the core ideas of complexity science, and argues convincingly that the basic laws, even in biology, will ultimately turn out to be simple. ■

Mark Buchanan is a freelance writer based in the United Kingdom.

Diversity with a difference

The Natural History of Madagascar

edited by Steven M. Goodman
& Jonathan P. Benstead

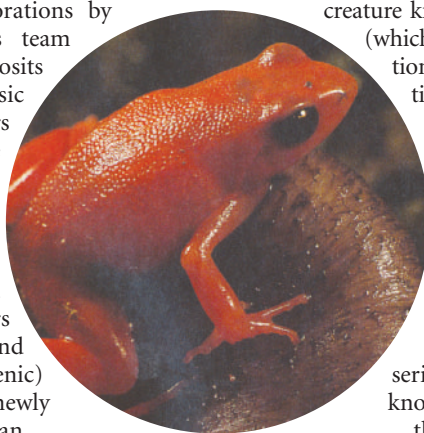
University of Chicago Press: 2004. 1,760 pp.
\$85, £59.50

Tim Flannery

Madagascar is perhaps the strangest place on Earth — an island where things can seem beguilingly familiar but on closer inspection take on an utterly alien aspect. Some of Madagascar's spiny insectivores known as tenrecs look like hedgehogs, but live very unhedgehog-like lives, anointing themselves with toxic chemicals and communicating by means of an organ of stridulation, from which eerie tones emerge, caused by the rattling of quills. Even evolution on the island seems to have occurred in the wrong order: elsewhere, rainforests commonly represent the ancestral (early Tertiary) environment, yet Madagascar's most venerable vegetation is its semi-arid, baobab-dominated scrubland.

Until 90 million years ago, Madagascar was part of the supercontinent of Gondwana, and its fossil record, much of it only recently revealed, is highly surprising. Palaeontological explorations by John Flynn and his team uncovered rich deposits dating to the mid-Jurassic (about 160 million years ago) that contain the remains of the world's oldest advanced (tribosphenic) mammal. And David Krause and co-workers discovered that, 70 million years ago, marsupials and archaic (non-tribosphenic) mammals shared the newly isolated island with an astonishing diversity of dinosaurs and crocodiles, including a herbivorous crocodilian with flattened teeth and a turtle-like skull.

All these creatures seem to have become extinct without leaving descendants on the island — the ancestors of Madagascar's lemurs and most other living mammals (as well as many flying invertebrates and birds) all arrived by sea during the past 60 million years. There are, however, some truly ancient endemics, whose ancestors have survived *in situ* since the age of dinosaurs — including reptiles such as the freshwater turtle *Erymnochelys madagascariensis*, whose nearest relatives are found in South America.



Unique to Madagascar: the Malagasy golden frog (*Mantella aurantiaca*).



The giraffe-necked weevil (*Trachelophorus giraffa*) is one of Madagascar's myriad bizarre creatures.

As exciting as Madagascar's biodiversity is today, the first humans to step ashore encountered an even more extraordinary world. The date is not definitely known, although it may have been after William the Conqueror crossed the English Channel in 1066. Amazingly, the settlers did not come from nearby Africa, but from southeast Asia. Those lucky pioneers would have seen, among other wonders, the world's largest bird (the elephant bird *Aepyornis maximus*), a lemur the size of a gorilla, a creature known as *Plesiorycteropus* (which is so odd that its relationships cannot be definitively established), giant tortoises and pygmy hippos. By the fourteenth century, most of these were extinct, and today the largest of Madagascar's native land-based animals weighs less than 10 kilograms.

For those who are serious about getting to know this fascinating island, there is no better resource than *The Natural History of Madagascar*. It is the closest thing to a comprehensive natural history of the region ever produced. Containing detailed accounts of flora, fauna, human ecology and the marine environment, even obscure groups such as the freshwater diatoms and springtails gain some coverage, although puzzlingly there is no chapter on the island's gymnosperms (pines and their relatives).

The extensive tabulated lists accompanying many of the systematic accounts will be invaluable to specialists, as will the exhaustive references that complete each major section. Such a book could have been dull, but this volume is rescued by the inherent

interest of creatures such as Dracula ants, whose queens subsist by sucking the blood of their larvae, and tales of scientific discovery, such as the finding in 1994 of the island's only species of Winteraceae (a family of primitive flowering plants), which was last seen in 1909.

The book also makes it clear that Madagascar's biodiversity is under severe threat. In their chapter on freshwater fishes, John Sparks and Melanie Stiassny note that neither had ever "witnessed anything surpassing the degree and extent of environmental degradation in Madagascar", and that "it is likely too late to save all but small remnants" of the island's unique fishes. This is particularly tragic given that Madagascar's fish fauna is highly endemic and that some of its fishes are ancient relatives of groups that today dominate elsewhere.

Despite such gloomy prognostications, *The Natural History of Madagascar* provides hope for a better future, as it warns of the disastrous losses to come and brims with innovative ideas about how to stave off the looming ecological catastrophe. The notion that seems to me to have particular merit is David Burney's suggestion that ostriches and the Aldabran giant tortoise (which may be the same species as Madagascar's extinct giant tortoise) be returned to the Madagascar grasslands, where they might help to control fire, offer a source of meat and act as a focus for tourism.

Yet the most encouraging thing about this book is surely its large number of Malagasy contributors, for it bespeaks a growing body of local expertise in biological sciences and conservation, without which any attempt to save the biodiversity of the most extraordinary place on Earth would be doomed.

Tim Flannery is director of the South Australian Museum, North Terrace, Adelaide, South Australia 5000, Australia.

The perils of anthropomorphism

Consciousness should be ascribed to animals only with extreme caution.

Clive D. L. Wynne

The complexity of animal behaviour naturally prompts us to use terms that are familiar from everyday descriptions of our own actions. Charles Darwin used mentalistic terms freely when describing, for example, pleasure and disappointment in dogs; the cunning of a cobra; and sympathy in crows. Darwin's careful anthropomorphism, when combined with meticulous description, provided a scientific basis for obvious resemblances between the behaviour and psychology of humans and other animals. It raised few objections.

The 1890s saw a strong reaction against ascribing conscious thoughts to animals. In the United Kingdom, Conwy Lloyd Morgan's canon forbade the explanation of animal behaviour with "a higher psychical faculty" than demanded by the data. In the United States, Edward Thorndike advocated replacing the use of anecdotes in the study of animal behaviour with controlled experiments. He argued that when studied in controlled and reproducible environments, animal behaviour revealed simple mechanical laws that made mentalistic explanations unnecessary.

This rejection of anthropomorphism was one of the few founding principles of behaviourism that survived the rise of ethological and cognitive approaches to studying animal behaviour. But after a century of silence, recent decades have seen a resurgence of anthropomorphism.

This movement was led by ethologist Donald Griffin, famous for his discovery of bat sonar. Griffin argued that the complexity of animal behaviour implies conscious beliefs and desires, and that an anthropomorphic explanation can be more parsimonious than one built solely on behavioural laws. Griffin postulated, "Insofar as animals have conscious experiences, this is a significant fact about their nature and their lives." Animal communication particularly impressed Griffin as implying animal consciousness.

Griffin has inspired several researchers to develop ways of making anthropomorphism into a constructive tool for understanding animal behaviour. Gordon Burghardt was keen to distinguish the impulse that prompts children to engage in conversations with the family dog (naïve anthropomorphism) from 'critical anthropomorphism', which uses the assumption of animal consciousness as a "heuristic method to formulate research agendas that result in publicly verifiable data that move our understanding of behaviour forward." Burghardt points to the death-feigning behaviour of snakes and possums as



Lost in thought: but do animals think like we do?

examples of complex and apparently deceitful behaviours that can best be understood by assuming that animals have conscious states.

Burghardt's critical and naïve anthropomorphisms are akin to Frans de Waal's distinction between 'animal-centred' and 'anthropocentric' anthropomorphism. For de Waal, denying human-like thought processes in cases where its assumption is warranted is just as dangerous as to anthropomorphize inappropriately. So seriously does de Waal view this error that he terms it 'anthropodenial' — "a blindness to the human-like characteristics of animals, or the animal-like characteristics of ourselves". De Waal proposed his research on reconciliation in chimpanzees as an example of the practical utility of anthropomorphism. Earlier studies had viewed aggression within a group as a means of maintaining space between individuals — which predicts reduced contact between individuals after aggressive encounters. De Waal used anthropomorphism to predict that the need for social cohesion would lead to reconciliation after conflicts — this was confirmed by an increased rate of affiliative contact between combatants after aggression.

But anthropomorphism is not a well-developed scientific system. On the contrary, its hypotheses are generally nothing more than informal folk psychology, and may be of no more use to the scientific psychologist than folk physics to a trained physicist. Although anthropomorphism may on occasion be a source of useful hypotheses about animal behaviour, acknowledging this does not concede the general utility of an anthropomorphic approach to animal behaviour.

John Staddon's model of animal short-term memory likens the process of forgetting to fluid leaking from buckets. But the utility of this model does not point to any general similarity between animal memory and fluid in buckets.

The philosopher Daniel Dennett argues along with Griffin, Burghardt and de Waal that appeals to mentalistic intentional states are more parsimonious than "an abstemiously behaviouristic story of unimaginable complexity". Certainly, attempts to explain complex behaviour in terms of simple mechanisms can become highly convoluted. But to argue that mentalistic accounts are simpler is, as Mark Blumberg and Ed Wasserman point out, to commit the nominal fallacy — to believe that giving something a name is the same as explaining it.

Consider the dance language of honeybees that impressed Griffin as possible evidence of consciousness. Foraging bees communicate the direction, distance and quality of a nectar source to their hive-mates. However, their 'language' is much more limited than human language. Honeybees communicate only three dimensions of experience, and only about two things (nectar sources and potential hive sites). Human communication is, in principle, unlimited in the number of dimensions of experience it can convey. To point out that if people were to communicate about nectar sources, we would do so consciously, offers no clue as to how bees might do it. Analogies at least as useful in understanding honeybee communication can be found in the activity of networked computers (unconsciously) establishing communication protocols.

Old-time behaviourism may have imposed excessive constraints on animal psychology. But the reintroduction of anthropomorphism risks bringing back the dirty bathwater as we rescue the baby. The study of animal behaviour is not so mature that it can thoughtlessly reject analogies from any source. But progress will surely be most rapid when we adopt explanatory frameworks that are concrete and unambiguous. Anthropomorphism, even critical or animal-centred, cannot offer that constructive basis. ■

Clive D. L. Wynne is in the Department of Psychology, University of Florida, Gainesville, Florida 32611-2250, USA.

FURTHER READING

Blumberg, M. S. & Wasserman, E. A. *Am. Psychol.* **50**, 133–144 (1995).

De Waal, F. B. M. *Phil. Top.* **27**, 255–280 (1999).

Mitchell, R. W. *et al.* *Anthropomorphism, Anecdotes and Animals* (State Univ. New York Press, New York, 1997).

Wynne, C. D. L. *Do Animals Think?* (Princeton Univ. Press, Princeton, New Jersey, 2004).

Lost in translation

Kenneth R. Chien

The potential use of stem cells as agents of repair in human disease makes them the subject of high-profile studies. But we should be wary of prematurely pushing laboratory research into clinical practice.

On the road map for 'translational medicine' — often referred to as bench-to-bedside research — stem-cell therapy is a prime destination. Stem cells are cells that have not taken on the identity of any specific cell type and are not yet committed to any dedicated function; they can divide indefinitely and may be induced to give rise to one or more specialized cell types. They not only offer scientists a tool to study the early molecular events in organ development, they also offer hope for tissue repair and regeneration to patients suffering from a spectrum of degenerative diseases.

Given this level of excitement, it is hardly surprising that the theory-to-therapy approach to the use of stem cells would be thrust forward into translational human studies at the earliest possible stage. A prompt for action came in 2001 with the publication of a paper in *Nature* by Orlic and colleagues¹ suggesting that stem cells derived from bone marrow can replace heart muscle lost as a result of heart attack, and can improve cardiac function. Injecting bone-marrow stem cells into an injured heart potentially represented a new therapy, triggering the launch of numerous clinical studies to investigate the effect of directly injecting these cells into the damaged heart muscle of patients following a heart attack. In two papers in this issue, independent studies by Murry *et al.*² and Balsam *et al.*³ seriously challenge Orlic and colleagues' initial observations and the scientific underpinnings of the ongoing human studies.

Both sets of authors used state-of-the-art genetic tools to examine whether bone-marrow stem cells transplanted into damaged hearts (Fig. 1) could take on the role of heart muscle cells and improve heart function. Murry *et al.*² used bone-marrow stem cells that included a foreign gene — *LacZ* — in their genetic complement and that only expressed it when they were in the heart muscle, creating a specific molecular tag. Little or no activity of this 'reporter' could be seen after the stem cells were injected directly into damaged hearts in mice. Balsam *et al.*³ found that bone-marrow stem cells labelled with a different tag — green fluorescent protein (GFP) — also showed little evidence of becoming cardiac muscle cells after they were transplanted into damaged mouse hearts. Not only that, the stem cells developed into different

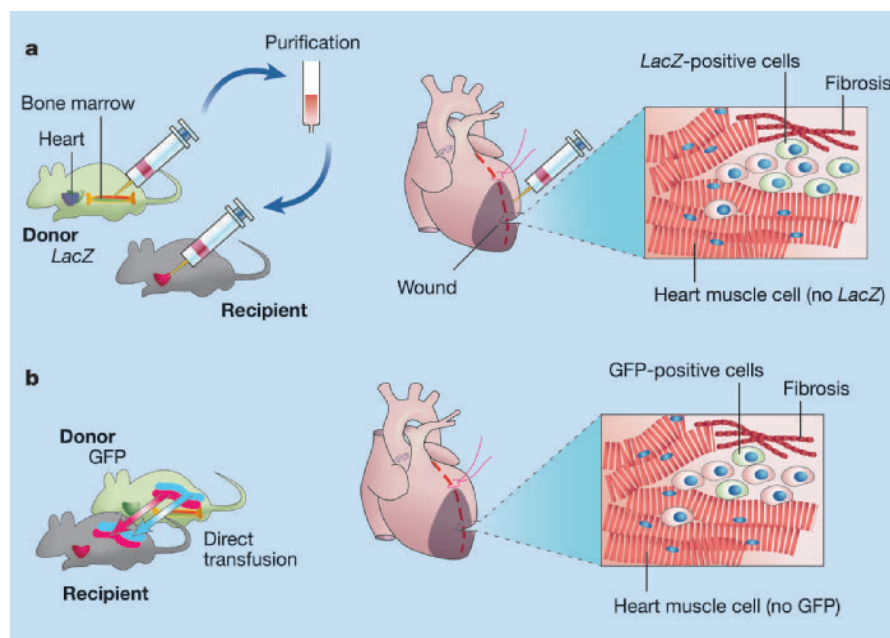


Figure 1 Two strategies used to show that bone-marrow stem cells do not take on the role of damaged heart cells. **a**, Murry *et al.*² isolated and purified genetically modified bone-marrow stem cells from mice. The modification 'tagged' the cells (with *LacZ*), enabling them to be detected in the recipient mouse heart, into which the cells were directly injected. Closer inspection of the recipient heart showed that the label could not be detected in heart muscle cells. **b**, Similar results were shown by Balsam *et al.*³, although the approach was slightly different. Donor bone-marrow stem cells were transfused directly into the circulation of recipients. Again, the tag (GFP; green fluorescent protein) could not be detected in heart muscle cells of the donor; indeed, the bone-marrow cells continued to differentiate into blood cells while in the heart.

blood-cell types, despite being in the heart.

Although Orlic *et al.*⁴ have also previously suggested that it is normal for bone-marrow stem cells to circulate in the blood, continually homing, repopulating and regenerating heart muscle cells, these two new studies indicate that this is probably not the case. GFP-tagged bone-marrow stem cells that were continually transfused into the bloodstream of a mouse with a large scar resulting from a heart attack showed no signs of becoming heart muscle cells³. The new reports raise serious concerns regarding the feasibility of using stem cells derived from the bone marrow to drive cardiac regeneration.

So scientists are asking why there are wide discrepancies between the earlier report and the current investigations. As Murry *et al.*² suggest, the differences may arise from the difficulty of tracking the *in vivo* fate of transplanted cells within an intact organ. Orlic *et al.*¹ mainly relied on detecting unique

protein constituents of bone-marrow stem cells using fluorescently tagged antibodies. Murry *et al.*² and Balsam *et al.*³, however, created intrinsic genetic markers that can be easily recognized without antibody staining. Owing to its high density of muscle-specific contractile proteins, intact heart muscle tends to have a high inherent background fluorescence, and can also display nonspecific antibody binding to the abundant muscle proteins. This makes it difficult, even for the most experienced labs using the most specialized microscopes, to track cell fate by simply using techniques that rely on fluorescent antibody staining of cardiac proteins.

For exactly this reason, genetic markers have long been used as the most reliable way to trace cell lineages during heart development (for example, see ref. 5). Similarly, specific genetic techniques that can determine whether stem cells have taken on muscle-cell-specific behaviour because they have fused with native muscle cells, rather

than because their genes have genuinely become reprogrammed to assume the new cell type (transdifferentiation), have become routine in studies of transplanted stem cells in the heart^{6,7} and many other organs^{6,8,9}.

Over the past few years, various experiments using several types of stem cells have supported the view that transdifferentiation occurs rarely, if at all, in many organ systems, including heart muscle^{6,7}. Less than 2% of the transplanted or injected cells take on the *in vivo* fate of heart cells. If this is true, then the improvement in cardiac function seen by Orlic *et al.*¹ might have arisen not because the stem cells transdifferentiated, but because new blood vessels were encouraged to grow around the injected area. Such growth of new blood vessels has been consistently found in transplantation studies of diverse cell types in the heart (for a review, see ref. 10). Recently, studies in large-animal models of transplanted bone-marrow-derived stem cells in the injured heart also failed to document cardiac regeneration. Again, the implication is that any functional improvement seen may not be related to an increase in functioning heart muscle *per se*¹¹. In addition, a recent clinical study, in which bone-marrow stem cells were transplanted into injured hearts, has been terminated because of serious cardiac side effects that threatened the blood flow to the heart¹². This again suggests that further experimental testing is warranted in large-animal model systems.

For physicians, the use of stem-cell therapy in treating cardiac-muscle diseases remains a worthy, but perhaps longer-term, goal. A glance at what formed the basis for the use of another cell type — myoblasts (progenitors of non-cardiac muscle cells) — in clinical trials for transplanting to damaged hearts is highly instructive (Fig. 2; for a review, see ref. 13). These clinical studies, which used patients' own myoblasts in the hope of increasing the amount of viable muscle after a heart attack, were the culmination of more than seven years of painstaking basic research. They showed that the transplanted myoblasts could adopt a muscle fate *in vivo* and improve heart function. A rigorous system was implemented to generate sufficient quantities of clinical-grade myoblasts from each patient to allow these cells to be directly injected into their own injured hearts.

The early signs were cautiously encouraging, although there were reports of the patients developing an abnormal heartbeat. This was probably due to electrical disorganization, caused by the failure of the transplanted cells to couple electrically with their neighbours to maintain synchronous contraction and impulse propagation during each heartbeat. Nonetheless, an overriding improvement in heart function was seen in the remaining patients. In short, we have had proof-of-concept that transplantation

- 1 Demonstration that the progenitor/stem cell of interest can differentiate *in vitro* into fully mature cardiac muscle cells in the absence of cell fusion.
- 2 Documentation that the cell type can function as an authentic *in vivo* cardiac progenitor in the embryonic or postnatal heart, and can be localized to a specific region of the intact heart.
- 3 Identification of specific, defined molecular cues that drive the differentiation of the progenitor/stem cells into physiologically functioning cardiac muscle cells.
- 4 Use of genetic markers to identify cells within an intact organ/tissue; re-isolation of these cells, and documentation of differentiation (molecular and physiological) that is indistinguishable from differentiated cells from the same preparation.
- 5 Improvement in organ/tissue function clearly ascribed to differentiation of the cell type of interest.
- 6 Documentation of stability, electrical coupling and lack of long-term loss of specialization or transformation events.
- 7 Direct demonstration that results in genetically modified mice can be extrapolated to large-animal models of heart failure that better reflect human cardiac physiology.

Figure 2 From experiment to therapy. A stepwise translational pathway for systematically moving experimental studies of cardiac stem cells into potential clinical therapy for heart failure.

of muscle-cell progenitors may improve function after a heart attack — the task is to find the ideal source of cells.

Previous results suggest that fetal heart cells might be the optimal cell type for heart-cell transplantation therapy¹⁰. They graft easily, adopt the identity of an adult cardiac cell type without fusion, and are electrically coupled. But the controversy that surrounds their use in scientific research means that they cannot be used in clinical studies, at least for the foreseeable future.

Perhaps now is the time to search for the

presence of naturally occurring, authentic heart progenitor cells, and to identify and dissect the signals that guide their migration, renewal and differentiation, as opposed to running the risk of becoming lost in translation with cell types that can rarely become functional heart muscle cells. This will require a rigorous approach to defining the molecular pathways that maintain the pool of endogenous cardiac progenitors during the process of heart formation itself. The question will then be whether remnants of these cells can be isolated from the fully developed adult heart and expanded in sufficient quantities to allow transplantation. In short, the key to unlocking the potential of cardiac stem-cell therapy may lie in a firm understanding of the biology of embryonic cardiac progenitors during development (see ref. 14 for a review).

For scientists, physicians and patients, the long-term future of cardiac stem-cell therapy has never been brighter. But we first need to commit the necessary time and resources to identify the best cardiac stem-cell story to translate.

Kenneth R. Chien is at the Institute of Molecular Medicine, University of California, San Diego, La Jolla, California 92093, USA.
e-mail: kchien@ucsd.edu

1. Orlic, D. *et al.* *Nature* **410**, 701–705 (2001).
2. Murry, C. *et al.* *Nature* **428**, 664–668 (2004).
3. Balsam, L. B. *et al.* *Nature* **428**, 668–673 (2004).
4. Orlic, D. *et al.* *Proc. Natl Acad. Sci. USA* **98**, 10344–10349 (2001).
5. Cai, C. L. *et al.* *Dev. Cell* **5**, 877–889 (2003).
6. Alvarez-Dolado, M. *et al.* *Nature* **425**, 968–973 (2003).
7. Oh, H. *et al.* *Proc. Natl Acad. Sci. USA* **100**, 12313–12318 (2003).
8. Vassilopoulos, G., Wang, P. R. & Russell, D. W. *Nature* **422**, 901–904 (2003).
9. Wang, X. *et al.* *Nature* **422**, 897–901 (2003).
10. Dowell, J. D., Rubart, M., Pasumarthi, K. B., Soonpaa, M. H. & Field, L. J. *Cardiovasc. Res.* **58**, 336–350 (2003).
11. Bel, A. *et al.* *Circulation* **108**, Suppl. 1, II247–II252 (2003).
12. Kang, H.-J. *et al.* *Lancet* **363**, 751–756 (2004).
13. Menasche, P. *Heart Fail. Rev.* **8**, 221–227 (2003).
14. Chien, K. R. & Olson, E. N. *Cell* **110**, 153–162 (2002).

This article was published online on 21 March 2004
(doi:10.1038/nature02500).

Earth science

Time of reversal

Ronald T. Merrill

At intervals, the polarity of the Earth's magnetic field reverses. The timescale for this reversal is unclear — in fact, it seems to depend on the latitude of the site from which the geological data are extracted.

At the Earth's surface, the planetary magnetic field can vary on timescales that range over more than 18 orders of magnitude — from less than a millisecond to more than 100 million years. The most dramatic of these changes are magnetic-field reversals, in which the geomagnetic poles swap hemispheres. Several hundred such reversals have been documented from geological records. They seem to occur

randomly in time, with the shortest interval between successive reversals being 20,000–30,000 years and the longest about 50 million years. It is 40 years since the reality of field reversals became widely accepted; by now one might expect that what happens to the magnetic field during a reversal would be understood.

This is not so, even for the most recent reversal, which occurred about 790,000 years

Cell biology

Competitive edge

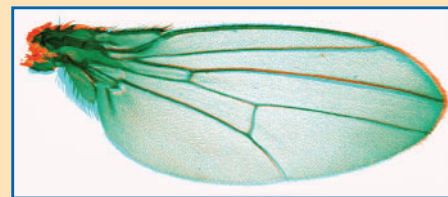
You might find it slightly unnerving that your cells sometimes compete fiercely with one another, but the process has been proposed to be a way in which organs control their size. Two groups have now taken a closer look at cell competition. Both describe what happens when levels of the growth-regulatory protein dMyc — the human counterpart of which is often hyperactive in cancer — are varied in developing fruitfly wings.

Laura A. Johnston and

colleagues (*Cell* **117**, 107–116; 2004) generated patches of cells that express high levels of dMyc, and found that these patches were much larger than controls, with bigger cells. Adjacent patches were smaller than usual, hinting that these cells were hampered by their faster-growing neighbours; in fact, the disadvantaged cells were killing themselves. This competition ensures that the excessive growth of cells with too much dMyc does not produce a bigger wing, as shown

here — a control wing (blue outline) is overlaid with one containing dMyc-overexpressing cells (orange).

Eduardo Moreno and Konrad Basler (*Cell* **117**, 117–129; 2004) came up with similar results. They suggest that excess dMyc prompts more protein synthesis and thus growth and proliferation, and that nearby cells — being smaller and less prolific — might be dying because they lose out in



the competition for external factors.

Whether cell competition controls organ size under normal circumstances remains to be seen. A further possibility — which prompted Moreno and Basler's study — is that a slight excess of growth regulator could result in adult cells ousting their normal neighbours and becoming cancerous.

Amanda Tromans

ago. It is not even known to what degree different reversals are similar to each other. On page 637 of this issue, Clement¹ sets out to improve on this state of affairs. He has extracted information on the structure and the duration of field transitions from several carefully chosen marine-sediment cores, and one finding in particular stands out: that the standard methods used to estimate the duration of a reversal produce different values, depending on the latitude of the site where the sediments were deposited.

The Earth's magnetic field is mostly generated by electric currents in the iron-rich liquid of the planet's outer core, at depths between 2,891 and 5,150 kilometres. A dynamo process converts the energy associated with fluid convection into magnetic-field energy. There are many models for the dynamo, including three-dimensional self-consistent models that exhibit magnetic-field reversals. But, because of computational difficulties, these models still use parameters that can be factors of a hundred or more different from those appropriate to the geodynamo problem². Consequently the model-builders look to palaeomagnetists to provide information on the properties of reversals that can be used to constrain the dynamo models — and to evaluate recent speculation that the Earth's field is at present in the initial stage of reversal³.

The sources of the magnetic field cannot be uniquely determined from measurements made at the Earth's surface. Instead, palaeomagnetists and geomagnetists often resort to a mathematical description of the magnetic field, known as spherical harmonics, that places all the field sources at the centre of the Earth. The most important component is the geocentric dipole, which can be mimicked by a bar magnet at the planetary centre. The extension of a line through the poles of this magnet intersects the surface of the Earth at two points, the north and south geomagnetic poles. If the Earth's magnetic field were a

pure dipolar field, the path of the north geomagnetic pole during a transition and the duration of the transition would be independent of the location of the site from which the data are taken.

Estimates based on this dipole-field assumption indicate that the field intensity decreases during a transition to a mean low value of about 25% of its usual value and that the duration of a transition is about 1,000 to 10,000 years⁴. Inconsistencies in these estimates have been attributed to a host of factors, including experimental and computational errors and the inadequacy of the dipole-field assumption. In fact, the present magnetic field can be divided into dipole and non-dipole parts, and it is expected that this is typical of fields in the past. A non-dipole field is characterized by a 'degree two', or higher, spherical harmonic, which in simple terms means that the non-dipole field has more than two locations where it is vertical at the surface. The total field at any location, at any time, is the vector sum of the dipole and non-dipole field.

For his analysis, Clement¹ has used drilled cores of marine sediment, carefully selected to contain magnetic minerals that would have accurately recorded reversals as the sediments formed into rock at the bottom of the oceans. The data from these cores indicate that the duration of the transition of the last reversal, 790,000 years ago, was about 2,000 years at the Equator, but about 10,000 years at mid-latitude (see Fig. 2 on page 637). It would be surprising if errors produced such a large systematic change, so Clement attributes this variation to the presence of a non-dipole field. Interestingly, Clement's data for three other reversals, although limited, suggest the same latitudinal variation. There might indeed be some common behaviour for the reversals that have occurred during the past few million years.

Clement¹ also presents simple reversal models in which a dipole, aligned with the

rotation axis, decays to zero and then builds up in the opposite direction, in the presence of a constant non-dipole field that is also aligned with the Earth's rotation axis. He argues that the model most compatible with the data has a constant, axial, octupole field — a degree-3, non-dipole field that is anti-symmetric about the Equator. This is somewhat unexpected, because other data and theory suggest that, in the middle of a transition, the field should mostly be symmetric about the Equator^{5,6}. However, as Clement's data come from discrete sites, his fit to the data cannot point to a unique model.

Moreover, this problem of uniqueness is exacerbated by the likelihood that the non-dipole field is also time dependent. Observational data and theory suggest that, usually, higher-degree fields (such as octupole fields) change faster than lower-degree ones (such as dipole fields)⁷. In addition, there are strong indications from dynamo theory that if the axial dipole field changes, so must the axial octupole field^{8,9}. So Clement's preferred model is not a unique solution, and it is also probably not physically realistic. But the work is an important contribution to our understanding of magnetic-field reversals on our planet, and it is unlikely to be the last paper on this subject to appear in the pages of *Nature*. ■

Ronald T. Merrill is in the Earth and Space Science Department, University of Washington, Seattle, Washington 98195-351310, USA.
e-mail: ron@ess.washington.edu

1. Clement, B. M. *Nature* **428**, 637–640 (2004).
2. Kono, M. & Roberts, P. *Rev. Geophys.* **40**, 1–53 (2002).
3. Hulot, G. et al. *Nature* **416**, 620–623 (2002).
4. Merrill, R. T. & McFadden, P. L. *Rev. Geophys.* **37**, 201–226 (1999).
5. Merrill, R. T. & McFadden, P. L. *J. Geophys. Res.* **93**, 11589–11597 (1988).
6. Li, J., Sato, T. & Kageyama, A. *Science* **295**, 1887–1890 (2002).
7. Christensen, U. R. & Olson, P. *Phys. Earth Planet. Inter.* **138**, 39–54 (2003).
8. Bullard, E. C. & Gellman, H. *Phil. Trans. R. Soc. Lond. A* **247**, 213–255 (1954).
9. Roberts, P. in *Mathematical Problems in the Geophysical Sciences* (ed. Reid, W. H.) 129–206 (Am. Math. Soc., Providence, Rhode Island, 1971).

Mars

Water, water everywhere

Timothy N. Titus

Mars is a very watery planet, but all the water seems to be frozen. Diving the amount and distribution of this water, past and present, is essential for understanding martian climates, and more.

For nearly 400 years, ever since Galileo first peered at Mars through a telescope, we have speculated about the existence of martian water and searched for it. Today, with three functioning spacecraft in orbit around the planet, and two robotic rovers on its surface, that search is more sophisticated than ever before. We can now be sure that water exists, and existed, on Mars. But our picture of its extent and distribution remains hazy.

The latest news comes from two different sources. The first is one of the rovers, Opportunity, which last month provided evidence of liquid water having once been at its landing site — a not-unexpected finding given that the site was selected from orbital data¹ showing the presence of grey haematite, a mineral that often forms in the presence of liquid water. The second source is the accompanying report by Bibring *et al.*² (page 627), who, from data provided by the OMEGA instrument aboard the European Space Agency's Mars Express orbiter, describe the discovery of water ice at Mars's southern perennial ice cap. This, too, is not a surprise. But it is the most direct evidence that this cap is primarily composed of water ice, partially covered with a thin veneer of CO₂ ice. Here I'll discuss the background to this second development, and the implications.

Before the arrival of thermal-sensing instruments placed in orbit around Mars, both the seasonal and the perennial polar caps were believed to consist of water ice. In 1969, however, thermal observations taken by Mariner 7 showed the seasonal caps to be composed of CO₂ ice, not water ice. During the late 1970s, the Viking orbiters, with the capability of measuring both temperature and atmospheric water vapour content, demonstrated that the northern perennial cap was composed of water ice, whereas the top layer of the southern perennial cap was composed of CO₂ ice. Although our view of the northern perennial cap as predominantly water ice has remained essentially unchanged, the question of whether the southern cap's lower layers might be composed of water ice remained unanswered.

The first real evidence³ of water ice in the southern cap region was found in 1984. This came from the analysis of measurements, made in 1969 by ground-based instruments monitoring millimetre wavelengths, that showed an increase in the presence of water vapour during the southern summer that

was consistent with the exposure of a large amount of water ice at the surface. Unfortunately, this seems to have been a unique occurrence, and so has not been independently verified.

Since 1984, however, there has been further indirect evidence that the southern cap is mainly water ice (Fig. 1). The Mars Orbiter Laser Altimeter aboard the Mars Global Surveyor (MGS) provided elevation measurements of the polar cap that, when combined with computer models incorporating the properties of CO₂ and water ices, indicated that the southern perennial cap could not maintain its current shape if it were composed of CO₂ ice⁴. Rather, the elevation profile is consistent with a cap composed of 'dirty' water ice — that is, ice mixed with dust from the martian surface. Also, data from the Mars Odyssey Gamma Ray Spectrometer (GRS) revealed⁵ a considerable amount of the element hydrogen buried in the first metre of the substrate, especially in the polar regions. The most likely source of that hydrogen is water ice.

Further developments emerged last year, with independent analyses^{6,7} of data from the Mars Odyssey Thermal Emission Imaging System (THEMIS) that produced indications of exposed water ice on and near the southern perennial cap. Combined THEMIS imaging with data from the MGS Thermal Emission Spectrometer identified thermophysical properties consistent with exposed water ice at the edge of the perennial cap⁶. And the conclusion drawn from a study involving THEMIS imaging and computer modelling was that the perennial cap is mainly water ice, with a thin covering of CO₂ ice⁷.

Finally, another analysis⁸ has used GRS neutron spectrometer data to place con-

straints on water and CO₂ ices on the southern cap. The measured neutron counts can be modelled either as a thin layer of CO₂ ice covering a water-ice cap, or as a thick layer of CO₂ ice covering about 55% of a water-ice cap. The actual distribution of water and CO₂ ices probably lies between these two extremes.

Now we have the latest results from the OMEGA instrument aboard Mars Express. Bibring *et al.*² have used near-infrared reflection spectroscopy to identify water ice at the surface of the southern perennial cap. The martian surface reflects sunlight in differing amounts at different wavelengths, depending on the surface composition. This allowed the OMEGA team to distinguish between dust and CO₂ and water ices. Their analysis even leaves open the possibility of the presence of a fourth component — clathrates, crystalline substances in which one CO₂ molecule is trapped inside a water lattice.

Detailed analysis of OMEGA's observations, combined with thermal, neutron and γ -ray data from other instruments, should place further constraints on the percentage of water ice that is exposed and how much CO₂ ice may be covering it. A vast quantity of water ice is also believed to be buried in the region surrounding the cap, known as the polar layered deposits, and Bibring *et al.* see outcroppings of exposed water ice on these deposits. All told, the surface water ice OMEGA has identified is probably just the tip of the iceberg.

Two implications are apparent in the new results. First, the southern region is indeed a major reservoir of water. Second, and perhaps more notable for understanding past climates on Mars, that region is not a significant reservoir of CO₂. Many researchers have suggested that during periods of high obliquity — when the axis of Mars was tilted more towards the Sun than it is today — the CO₂ in the southern perennial cap would be released, increasing the pressure of the planet's atmosphere. The argument is that this increase in atmospheric pressure might have been sufficient to allow liquid water to become stable on the surface. Although dust and rubble on the martian surface may still contain an appreciable amount of CO₂ that

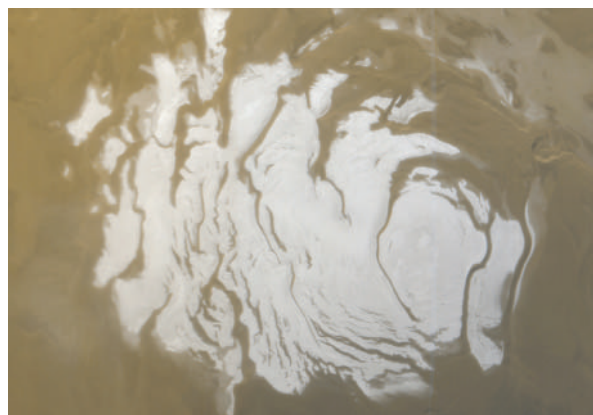


Figure 1 The south polar cap of Mars, imaged in April 2000 by the Mars Orbiter Camera aboard the Mars Global Surveyor. As shown here, in the martian southern summer, the cap is about 420 km across. This is its smallest extent; from June, the area covered by frost grows until it fills the area shown in this image. The surface area of the southern perennial cap has been estimated⁹ as $1.44 \times 10^6 \text{ km}^2$, with a volume of $2\text{--}3 \times 10^6 \text{ km}^3$.

could be released, the OMEGA results described by Bibring *et al.* exclude any possibility that the southern perennial cap could appreciably affect the atmospheric pressure.

The search for the existence of water on Mars, past and present, will continue. Future reports will often confirm less-direct evidence, as is the case with the information about the widespread existence of water ice at the martian south pole provided by OMEGA. But this does not lessen the importance of these discoveries.

Did Mars ever support life? Will Mars support human life in the future? The answers depend on understanding the past and present distribution of both water and CO₂. Life, as we know it, requires liquid water. Yet the long-term stability of liquid surface water requires a thicker atmosphere than Mars has at present. OMEGA's observations show that its past atmosphere, predominantly CO₂, is not locked up in the polar caps. We can hope that OMEGA and her sister instruments on Mars Express, Mars Global Surveyor and Mars Odyssey, along with reports from the

rovers Spirit and Opportunity, will deliver the necessary information to tell us where that ancient atmosphere went.

Human exploration and ultimate colonization of Mars depend on accessibility of one resource — water. Martian water is necessary not only for human consumption, but is also the key to making breathable air and fuel for a return trip to Earth. For life on Mars, water is the elixir. ■

Timothy N. Titus is at the US Geological Survey, 2255 North Gemini Drive, Flagstaff, Arizona 86001, USA.

e-mail: ttitus@usgs.gov

1. Christensen, P. R. *et al.* *J. Geophys. Res.* **105**, 9623–9642 (2000).
2. Bibring, J.-P. *et al.* *Nature* **428**, 627–630 (2004).
3. Jakosky, B. M. & Barker, E. S. *Icarus* **57**, 322–334 (1984).
4. Nye, J. F. *et al.* *Icarus* **144**, 449–455 (2000).
5. Boynton, W. V. *et al.* *Science* **297**, 81–85 (2002).
6. Titus, T. N., Kieffer, H. H. & Christensen, P. R. *Science* **299**, 1048–1051 (2003).
7. Byrne, S. & Ingersoll, A. *Science* **299**, 1051–1053 (2003).
8. Prettyman, T. H. *et al.* *J. Geophys. Res.* (in the press).
9. Smith, D. E. *et al.* *Science* **284**, 1495–1503 (1999).

This article was published online on 17 March 2004 (doi:10.1038/nature02482).

Behavioural evolution

Cooperate with thy neighbour?

Peter D. Taylor and Troy Day

What gives cooperation an evolutionary edge? Two features of a population — spatial structure and finite size — are factors in the success of any strategy, although more subtle than we thought.

In thinking about the evolution of cooperative behaviour¹, there is one main stumbling block: that cooperative individuals can be exploited by 'defectors', who benefit from cooperation while avoiding the costs that it entails. Solutions to this problem typically find ways for cooperative individuals to interact with one another more often than they would purely by chance. There are two basic ways in which this can happen. One involves the population having a spatial structure with local reproduction and dispersal, so that neighbours of a cooperative individual are themselves more likely to be cooperative^{2,3}. The other relies on some form of information transfer whereby players can assess the behaviour of a prospective partner and decide accordingly how, or even whether, to play. The assessment might be made on the basis of traits that are reliable indicators of likely behaviour^{4,5} or through a phase of negotiation^{6,7}.

Two papers in this issue^{8,9} add further insight. Hauert and Doebeli⁸ (page 643) propose that, under certain conditions, spatial structure might actually hinder cooperative behaviour. It has long been understood that population structure can be a mixed blessing for cooperation, because the gains that it provides through positive assortment are

countered by competition between like individuals^{2,10,11}. Hauert and Doebeli have uncovered yet another limitation of population structure, one that also gives a fascinating geometric distinction between games such as Hawk–Dove — in this case, in the guise of the snowdrift game — and the Prisoner's Dilemma, or blizzard game (Box 1, overleaf).

For a spatially structured population of players, with their choice of strategy displayed as a particular colour on a grid, Hauert and Doebeli see a shift in the geometry of clusters of cooperators at the point where the cost and benefit of the encounter are equal. In the Prisoner's Dilemma, when the cost is greater than the benefit, globular clusters form (Fig. 1a), which give cooperators enough protection to persist at a small frequency. But in playing the snowdrift game, when benefit outweighs cost, the clusters become more finger-like, or dendritic (Fig. 1b). Here the cooperators are vulnerable to exploitation and they die out. The transition is perplexing, but it is clear that spatial structure in a population might not always work in favour of cooperation.

In the second article, Nowak *et al.*⁹ (page 646) suggest that finite population size is also crucial in the evolution of cooperation. These authors focus on the Prisoner's Dilemma



100 YEARS AGO

A New Mineral from Ceylon. In the beginning of February I bought from Mr. Holland 5 cwt. of the mineral described by Prof. Dunstan in last week's *NATURE* (p. 510)... I had hoped to have positive and definite results to communicate before describing its constituents, but the publication by Prof. Dunstan of an analysis, and his statement that he is still engaged in its investigation, makes it necessary to write this letter... Fractionation shows that the oxalate precipitate (the portion soluble in ammonium oxalate) gives equivalents between 25.0 (the most insoluble portion of the double sulphate) and 44.7 (the most soluble portion); by far the major part of the element has the last mentioned equivalent... Assuming that the element is a tetrad, which is probable from its behaviour, it undoubtedly possesses an equivalent approaching the highest number (44.7), and for this there is a gap in the periodic table between cerium and thorium; one at least of the elements present (supposing that there is more than one present) will probably have an atomic weight of about 177, preceding tantalum (182.5) in the horizontal row of the periodic table... Within the limits of this letter I am obliged to omit many more characteristics of this curious ore... I regret to have been obliged to tell an imperfect story. William Ramsay
From *Nature* 7 April 1904.

50 YEARS AGO

The brain mechanisms which serve the sun navigation of the animals mentioned are presumably of a similar nature in all species and are probably based on the same principles as human sun orientation. They can also be expected to share certain properties with other time-keeping mechanisms (internal clocks) which are of wide occurrence. They keep time fairly well on their own and they are set and kept in pace by light stimuli. Their metabolic nature has in certain cases been established. A third property of such systems is that they regulate motor activities. *Drosophila*, for example, normally emerge from their pupæ before dawn. If a bottle with larvæ and pupæ of *D. melanogaster* is artificially illuminated during three consecutive nights and kept in darkness during daytime, the flies which emerge during the following week will 'remember' the time of the artificial dawn and emerge in the evenings, even when now kept in perpetual darkness. H. Kalmus
From *Nature* 10 April 1954.

to highlight dramatically the difference between evolutionary stability in a finite and an infinite population, and at the same time suggest a new factor that bears on the evolution of cooperation.

In the Prisoner's Dilemma, defectors always outcompete cooperative individuals when encounters are random. Axelrod and Hamilton demonstrated¹², however, that cooperative strategies can be enhanced if multiple encounters with the same partner are allowed and if current behaviour is based on past experience. Of all such conditional strategies, 'tit-for-tat' seems to be one of the best: a player cooperates initially but continues to cooperate only if its partner cooperated in the previous encounter. It turns out that if the number of encounters with the same partner is large enough, tit-for-tat can outperform a uniform all-defect strategy once its frequency is high enough. This means that there is an unstable mixed equilibrium at some particular frequency: above it, tit-for-tat dominates; below it, all-defect takes over. At least, this is true of an infinite population (in which changes in frequency are deterministic in evolutionary time). For example, with a benefit of 3, a cost of 4, and 10 encounters per partner, this unstable equilibrium is at a frequency of 1/8 tit-for-tat.

But what if the population is finite, say of size 80? Treated as an infinite population, we'd expect to need ten individuals playing tit-for-tat before this strategy could become more fit than all-defect; and by the standard definition, all-defect is evolutionarily stable (no rare mutant tit-for-tat-er can invade). In contrast, as a finite population, the stochastic nature of random sampling leads us to expect that, after many generations, all individuals will be descended from exactly one of the original members. With neutral strategies, each individual would have the same

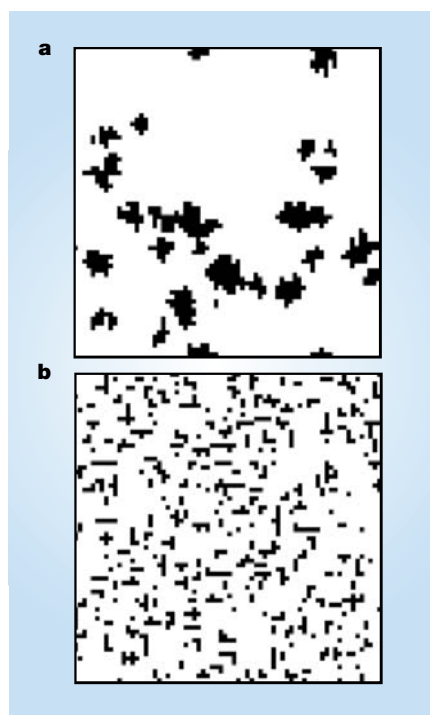


Figure 1 Cooperators versus defectors. With the spatial structure of a population represented by a grid, Hauert and Doebeli⁸ find different equilibrium configurations for two slightly different games that test the evolution of cooperation. Cooperators are shown in black, defectors in white. **a**, In the Prisoner's Dilemma game, clusters of cooperators develop and can offer protection to those in the interior of each cluster, increasing the fitness of cooperators. **b**, In the snowdrift game, however, the cooperative clusters develop into dendritic fingers that poke out into defector territory, exposing their members to exploitation. Cooperators can actually be worse off than if they had formed partnerships at random.

probability of being the founder, which suggests an alternative way of comparing the fitness of tit-for-tat versus all-defect — calculate the probability that an individual of each kind will be the founder¹³.

It turns out that, for the example of one lone tit-for-tat-er in a population of 79 all-defect players, the probability that the tit-for-tat individual is the founder is almost twice that of an all-defect individual (M. Nowak, personal communication). Should we still regard all-defect, then, as an evolutionarily stable strategy? In fact, Nowak *et al.*⁹ use this example to propose an extension of the standard definition of evolutionary stability for finite populations. The mutant strategy must be less fit in two ways: no rare mutant can invade (the traditional sense), and a rare mutant individual must have a lower than normal chance of being the founder of the ultimate population. Certainly, in a finite population such an extension is needed, but it's not at first so clear how to do this, nor exactly what role fixation probability should play. In a population subject to the forces of mutation and drift, what are the states that we might expect to observe?

Peter D. Taylor and Troy Day are in the Department of Mathematics and Statistics, Queen's University, Kingston, Ontario K7L 3N6, Canada.

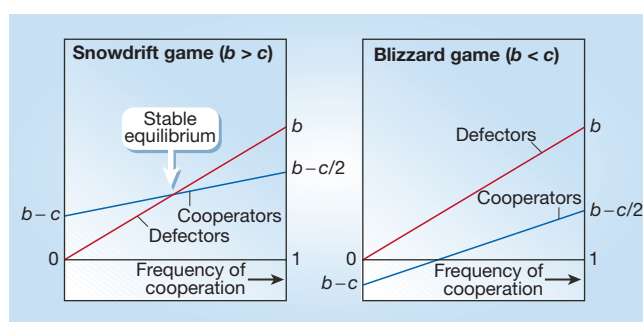
e-mail: taylorp@post.queensu.ca

1. Dugatkin, L. A. *Cooperation Among Animals* (Oxford Univ. Press, 1997).
2. Hamilton, W. D. in *Man and Beast: Comparative Social Behavior* (eds Eisenberg, J. F. & Dillon, W. S.) 59–91 (Smithsonian Inst. Press, Washington, 1971).
3. Wilson, D. S. *Proc. Natl Acad. Sci. USA* **72**, 143–146 (1975).
4. Nowak, M. A. & Sigmund, K. *Nature* **393**, 573–577 (1998).
5. Riolo, R. L. *et al. Nature* **414**, 441–443 (2001).
6. McNamara, J. M. *et al. Nature* **401**, 368–371 (1999).
7. Taylor, P. D. & Day, T. *Proc. R. Soc. Lond. B* **271**, 669–674 (2004).
8. Hauert, C. & Doebeli, M. *Nature* **428**, 643–646 (2004).
9. Nowak, M. A. *et al. Nature* **428**, 646–650 (2004).
10. Wilson, D. S. *et al. Evol. Ecol.* **6**, 331–341 (1992).
11. Taylor, P. D. *Proc. R. Soc. Lond. B* **249**, 299–302 (1992).
12. Axelrod, R. & Hamilton, W. D. *Science* **211**, 1390–1396 (1981).
13. Proulx, S. R. & Day, T. *Selection* **2**, 1–16 (2001).

Box 1 Snowdrifts and blizzards

Imagine two car drivers caught in a snowdrift. The total cost of shovelling out of the drift is c and the benefit to each of getting home is b . The drivers might follow either of two strategies — cooperate with the other, or defect. If both drivers cooperate, they split the cost of shovelling and both get home. If one cooperates and one defects, they will both get home, but the cooperator will bear the whole cost. Two defectors bear no cost but get no benefit.

It is reasonable to assume that $b > c/2$ — that getting home is worth more than the cost of half the shovelling (or the game is pointless). That leaves two interesting cases: $b > c$ and $b < c$. The first case is



known as the snowdrift game and (in the spirit of comparison) we call the second the blizzard game. In the latter, the shovelling is so hard that a driver who does it all suffers a net loss. The snowdrift game is a version of Hawk–Dove, and the blizzard

game is a version of the Prisoner's Dilemma, both of which are much studied in behavioural evolution.

In a biological population in which the payoff contributes to fitness, we are interested in comparing the average fitness

of a cooperator and a defector. Fitness is illustrated here as a function of frequency of cooperative encounters. There is a stable equilibrium where the lines intersect. For random encounters between drivers/players, the snowdrift game supports a stable mixture of cooperators and defectors (roughly half-and-half). The blizzard game does not: the only point of stable equilibrium is an all-defector population. But if there were some mechanism that increased the frequency of cooperative encounters, the lower portion of the cooperation line would rise, creating a point of stable equilibrium for the blizzard game as well. **P.D.T. & T.D.**

Obituary

David Shoenberg (1911–2004)

David Shoenberg was the last survivor of the distinguished group of scientists who established low-temperature physics as a flourishing discipline in Britain before the Second World War. His father was the electronic engineer Isaac (later Sir Isaac) Shoenberg, who came to England from Russia with his family in 1914, and who led the team at EMI that developed television for the BBC in the 1930s.

At Cambridge in 1932, David Shoenberg became one of the very few research students taken on by Pyotr Kapitza, the brilliant Russian engineer–physicist who was then developing techniques for producing high magnetic fields. Kapitza set him on measuring the magnetostrictive effect — the minute length changes produced by a magnetic field — in bismuth. This was a project to test the ablest experimentalist, but one that Shoenberg completed successfully three years later. By that time he was on his own: in 1934, Kapitza had not been allowed to return from Russia after his annual summer holiday there. Shoenberg spent the next few years working mainly on superconductivity and writing a book that long remained the best introduction to the topic.

In 1937 Kapitza invited him to Moscow to work for a year in his newly built laboratory. There Shoenberg began to study in earnest the phenomenon that was to occupy him, very fruitfully, for the next 50 years. The de Haas–van Alphen (dHvA) effect is an oscillatory variation of a metal's magnetic susceptibility with magnetic-field strength at low temperatures, first observed in bismuth in 1930. Shoenberg and M. Z. Uddin had published a brief study of it in 1936, but now Shoenberg's measurements were far more sensitive and precise. It remained to interpret these strange oscillations.

Rudolf Peierls had already produced a rather cumbersome theoretical explanation, but fortunately the brilliant Russian theoretician Lev Landau was also working in Kapitza's laboratory, and he at once came up with a much more powerful formulation, making the analysis of the experimental results far easier. In 1938, however, Landau was declared an 'enemy of the people', and it became impossible to cite his work. All references to his theory were deleted when Shoenberg's work was published in a Russian journal, but Shoenberg was able to include a full account of it, correctly attributed, in the Royal Society version of his paper.



Pioneer in 'fermiology' — the study of electrons in metals

Since the advent of quantum mechanics in the 1920s, the field of solid-state physics had developed rapidly, and one of the central problems was to understand the properties of conduction electrons in metals. Considerable theoretical progress had been made, marked by the appearance of Hans Bethe's monumental article in the *Handbuch der Physik* in 1933, but so far little light had been shed on their behaviour experimentally.

Using Landau's theory (as generalized in 1951 by Lars Onsager and Ilya Lifshitz), Shoenberg made the dHvA effect into a powerful tool for understanding the behaviour of conduction electrons in metals. The electronic properties of metals are largely determined by the nature of the 'Fermi surface', a surface defined in momentum space that separates the lower-energy, filled electron states from the empty, higher-energy states. The science of 'fermiology' — the determination of the shape of this surface, and the velocities and collision rates of the electrons on it — has occupied many solid-state physicists profitably for many years, using the techniques developed by Shoenberg.

Bismuth was the first metal (or semi-metal) in which the oscillatory phenomenon had been observed, because it has a low density of conduction electrons (so its Fermi surface is unusually small), and the theory showed that this made the effect easier to observe in reasonably low magnetic fields. Until 1947 it remained the only material in which the dHvA effect had been observed, but in the next few years it was seen and studied in a variety of other

metals, by Shoenberg and others.

It proved remarkably difficult to observe the effect, however, in the noble metals copper, silver and gold, which required much higher magnetic fields. Shoenberg eventually solved the problem, elegantly and ingeniously, by discharging a bank of capacitors through a magnet coil to produce a field that rose rapidly to a peak and then fell again to zero. This field produced a time-dependent oscillatory magnetization in the sample, detected by a search coil wound around it. In 1958 he at last succeeded in detecting the dHvA effect in a copper whisker — although, as he wrote later in his definitive book on magnetic oscillations, "the breakthrough came by a combination of luck and slightly faulty reasoning". Such disarming honesty was characteristic of the man.

The pulsed-field apparatus was not easy to use. It was usually operated by Shoenberg's assistant E. Laurmann, a dour Estonian who had worked with Kapitza. On one occasion, when Laurmann went off for lunch while results were coming fast, Shoenberg tried his hand at operating the apparatus himself — the resultant bang when he accidentally short-circuited the capacitor bank left him permanently a little hard of hearing.

Since those days, the advent of superconducting magnets has made it much easier to attain the high fields needed in dHvA studies. The dHvA effect has been used extensively to study a variety of metals and many complex phenomena, including magnetic breakdown and the 'B–H effect', both discovered by Shoenberg and his students. Later studies include the 'Fermi-liquid effects' arising from electron–electron interactions, heavy-fermion effects, in which electrons seem to have a far heavier effective mass than simple theory would predict, and the 'mixed state' of superconductors. In all of these areas the dHvA effect has proved a powerful tool.

Until late in life, Shoenberg loved to travel abroad — to Russia, India, the United States and elsewhere — to meet friends and former students. Friendly, amiable and completely straightforward, an expert at asking the seemingly innocent but important question, David had a host of friends in many countries, who will miss him greatly.

R. G. Chambers

R. G. Chambers is Emeritus Professor of Physics, H. H. Wills Physics Laboratory, University of Bristol, Bristol BS8 1TL, UK.
e-mail: r.g.chambers@btinternet.com

Materials science

Spiders forced to spin an unusual yarn

Biomacromolecules doi:10.1021/bm034269x (2004)

Dragline silk, the fibre from which spiders make the scaffolding of their webs, is nature's undisputed high-performance fibre, and unravelling its secrets may help in developing ecologically friendly superfibres. But the forced 'silking' commonly used to obtain silk for experiments may seriously affect the fibres' properties, Christine S. Orllepp and John M. Gosline warn.

Filming the master spinners in action reveals that abseiling spiders initially fall freely and then slow down by applying friction to their dragline; the data show that the tension on the newly spun silk varies from less than 2% of the thread's breaking load during freefall to about 40% during the slowing-down phase. In contrast, forcibly pulling silk from strapped-down spiders, as occurs in silking, meets with active resistance — the spiders spin at much reduced speeds and apply braking forces that load the silk at up to 60% of its breaking point.

The prolonged braking during forced silking generates local heating. Not only may this induce spiders to abandon silk production, the authors argue, but it may also change the silk's structure. Further, the tension on newly spun silk should affect protein alignment within the thread before it dries, thus also altering its material properties.

Magdalena Helmer

Cell biology

Yeast chill out

Mol. Cell **13**, 771–781 (2004)

How do yeast cells stay alive when the temperature plummets to zero? By producing an armoury of protective molecules, according to Olga Kandror and colleagues.

The authors have examined the molecules made by budding yeast (*Saccharomyces cerevisiae*) as the temperature drops to near freezing. They find that, below 10 °C, the cells start to produce enzymes that synthesize large amounts of trehalose — a sugar known to protect organisms against high temperatures and to shield the bacterium *Escherichia coli* from the effects of a sudden drop in temperature. The yeast also generate a set of so-called heat-shock proteins (HSPs), which probably — as they do at high temperatures — bind to damaged proteins and prevent them from unfolding.

To their surprise, Kandror *et al.* also discovered that the hardy cells go on making these proteins and generating trehalose even at 0 °C. New messenger RNAs encoding HSPs and trehalose-synthesizing enzymes are made even though the production of



other mRNAs and proteins ceases. The authors also tested yeast that lack trehalose or the gene-transcription factors that make the necessary mRNAs: as expected, these mutants succumbed far more quickly in subzero conditions.

Michael Hopkin

Vertebrate evolution

Out on a limb

Science **304**, 90–93 (2004)

The advent of limbs for walking on land, and so of the tetrapods, was one of the main events in life's history. The traditional idea is that limbs arose from the fins of fish that were already living, at least in part, on land. But there are indications from the fossil record for the aquatic transition of fins to limbs that had the function of propping up the body in shallow waters.

Neil H. Shubin *et al.* now add to that evidence with their interpretation of a fossil limb bone, a humerus from a creature that lived about 370 million years ago. The authors compared the humerus with the equivalent bones from tetrapod and fish fossils of the same era. They point to its unique features, which, they say, represent an intermediate between fins designed as swimming aids and limbs for walking on land or in water.

Shubin *et al.* conclude that the humerus was not necessarily a component of a limb adapted for locomotion on land. But they argue that important transitions were evidently occurring in fish before the origin of the tetrapod limb with its characteristic elaboration of digits.

David Kramer

Nanoscale physics

Pure boron tubes

J. Phys. Chem. B **108**, 3967–3969 (2004)

Researchers have already made pure boron nanowires, and theoretical studies indicate that hollow boron nanotubes should be stable; they might even be better conductors of electricity than carbon nanotubes.

Dragos Ciuparu *et al.* now claim to have made the first pure boron nanotubes, by reacting boron trichloride with hydrogen

gas at 870 °C. The tubes are single-walled and measure about 3 nanometres in diameter. The reaction relies on a magnesium catalyst loaded onto mesoporous silica. The pores in the silica act as a template for the nanotubes, which grow up to 20 nm out of the catalyst surface.

The authors report that exposure to an electron beam damages the nanotubes, so electron-microscope images of them, taken at low dose, are rather fuzzy. However, the Raman infrared spectrum of the sample is consistent with a tubular structure. A more detailed structural analysis, omitting the electron beam, will follow.

Mark Peplow

Nanotribology

Slippery secrets revealed

Phys. Rev. Lett. **92**, 126101 (2004)

Explaining graphite's lubricating properties is rather more slippery than conventional wisdom suggests. The usual story is simply that graphitic carbon sheets slide over one another; but Martin Dienwiebel and colleagues show that a phenomenon called 'superlubricity' governs this process.

First proposed in 1990, the effect depends on whether the hexagonal arrays of carbon atoms in each sheet are in registry with one another. Sliding is relatively easy when there is no registry, but if the sheets are rotated until their crystallographic axes line up, they can lock together, with an abrupt increase in lateral friction. That, at least, was the theory. But only now is there good evidence for it.

Dienwiebel *et al.* have designed a new friction-force microscope in which a tungsten tip can move on 'soft' springs in any horizontal direction while being more stiffly confined vertically. The frictional force measured on graphite shows sawtooth-shaped profiles — 'stick-slip' behaviour — in all directions, but the magnitude of the oscillations peaks for substrate orientations 60° apart. This is what the superlubricity hypothesis would predict if, as the researchers think, a flake of graphite becomes attached to the probe tip, so that movements of the tip slide the flake over the substrate.

Philip Ball

Hummingbird jaw bends to aid insect capture

This tiny bird has a neat trick to trap flies in mid-air with its long nectar-seeking beak.

The upper jaws of birds, unlike those in many tetrapods, move relative to the skull and are often flexible along their length, whereas the lower jaw (mandible) is usually a rigid structure formed by the fusion of several bones, flexing only where it meets the skull. Here we describe a previously unnoticed mandibular bending movement in hummingbirds, in which the distal half of the mandible is actively flexed downwards and the gape widens to catch flying insects. The hummingbird is thought to have developed a long narrow bill as it specialized in feeding on floral nectar, but the bird's need to supplement its diet with insects must have contributed to the surprising flexibility of its jaw.

Hummingbirds (Trochilidae) have distinct specializations for nectar feeding¹, but they must also eat insects because floral nectars are deficient in essential amino acids^{2,3}. However, it is not clear how these birds adapt their feeding mechanics to meet the demands that result from consuming both nectar and insects.

The dissimilarity in beak shape between hummingbirds and birds dedicated to aerial insectivory is striking. Birds that only eat flying insects typically have short beaks with wide gapes, which provide a broad surface close to the mouth for efficient capture and transport of prey^{4,5}. By contrast, hummingbirds have long, narrow beaks that vary mainly in length and curvature.

We used high-speed video to study ruby-throated hummingbirds (*Archilochus colubris*), magnificent hummingbirds (*Eugenes fulgens*) and blue-throated hummingbirds (*Lampornis clemenciae*) as they hunted fruitflies (*Drosophila*, species unspecified). These hummingbirds catch their prey by flying at it with open jaws. During jaw opening, the distal half of the lower jaw flexes downwards (Fig. 1a). As the bird approaches prey, its jaw may bend without opening further, suggesting that bending is an active process and not a by-product of jaw opening (for movie, see supplementary information).

An intramandibular joint is found in other birds, such as the goatsuckers (Caprimulgidae), that feed exclusively on flying insects. In those cases, the joint flexes laterally, so that the lower jaw is bowed sideways by the opening of the mouth^{6–8}, and this enlarges the oral cavity. Presumably, this aids the capture of tiny prey. Dorsoventral flexion in the hummingbird mandible is mechanically and functionally linked to lateral jaw bowing. Our video recordings of hummingbirds reveal that flexion is

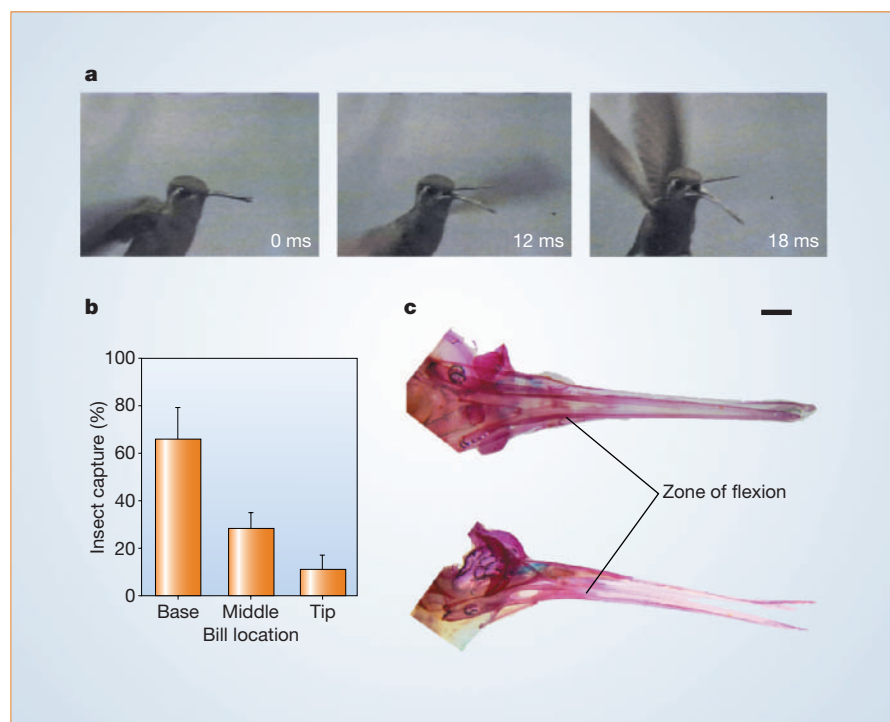


Figure 1 Intramandibular flexion during insectivory in hummingbirds. **a**, High-speed video sequence of an attack on a fruitfly (*Drosophila*; dark spot to the right) by a female magnificent hummingbird, showing the jaw opening, followed by bowing of the mandible at the base to widen the gape, and flexion of the distal half of the mandible. **b**, Success of insect capture (percentage caught) by a ruby-throated hummingbird in relation to where the prey first strikes the bill. More flies are caught at the flared base of the beak (repeated measures ANOVA; $F_{1,3} = 10.71$, $P = 0.047$); the few prey caught at the tip of the bill are often lost during transport to the mouth. **c**, Ventral (top) and right lateral (bottom) view of a cleared and stained (alizarin red, bone; Alcian blue, cartilage) ruby-throated hummingbird skull. No jaw joints are present; the mandible is a single, fused unit and has to deform in at least two dimensions in order to bend as seen in **a**. The intensity of the stain shows that the mandible is densely calcified up to the inflection point of dorsoventral bending. Scale bar: 1 mm.

accompanied by a bowing of the jaw that further widens the gape.

This intramandibular flexion is used by the hummingbird only during insect capture. Most successful captures occur at the gape, rather than at the tip or middle of the bill (Fig. 1b). By widening the gape, mandibular flexion probably makes it easier for prey to enter the mouth and decreases the transport distance while increasing the beak's effective capture surface.

We know of no other tetrapod in which intramandibular flexion is achieved in two dimensions simultaneously, is actively controlled, and is evident during feeding. This flexion in hummingbirds is particularly remarkable because there is no detectable intramandibular joint in the hummingbirds we examined: all mandibular elements are fused (Fig. 1c). Flexion in the two directions simultaneously must therefore require complex deformation of thin bone.

This previously undiscovered ability in a bird that has been as intensively studied as the hummingbird suggests that the focus on adaptations for nectar-feeding may

have obscured other important aspects of their evolution. Our results indicate that insectivory exerts a powerful influence on hummingbird form and function, despite their close evolutionary relationship with nectar-producing flowers.

Gregor M. Yanega, Margaret A. Rubega

Department of Ecology and Evolutionary Biology,
University of Connecticut, Storrs,
Connecticut 06269, USA

e-mail: gregor.yanega@uconn.edu

- Altshuler, D. L. & Clark, C. J. *Science* **300**, 588–589 (2003).
- Baker, H. G. & Baker, I. in *Biochemical Aspects of Evolutionary Biology* (ed. Nitecki, M. H.) 131–171 (Chicago Univ. Press, Chicago, 1982).
- Brice, A. T. & Grau, R. G. *Physiol. Zool.* **64**, 611–626 (1991).
- Storer, R. W. in *Avian Biology* (eds Farner, D. S. & King, J. R.) 150–188 (Academic, New York, 1971).
- Fitzpatrick, J. W. in *Ornithological Monographs* 447–470 (American Ornithologists' Union, 1983).
- Bühler, P. Z. *Morphol. Tiere* **66**, 337–399 (1970).
- Bühler, P. in *Form and Function in Birds* (eds King, A. S. & McLelland, J.) 439–468 (Academic, London, 1981).
- Zusi, R. L. in *The Skull: Patterns of Structural and Systematic Diversity* (eds Hanken, J. & Hall, B. K.) 391–437 (Chicago Univ. Press, Chicago, 1993).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.

Climatology

Threatened loss of the Greenland ice-sheet

The Greenland ice-sheet would melt faster in a warmer climate and is likely to be eliminated — except for residual glaciers in the mountains — if the annual average temperature in Greenland increases by more than about 3 °C. This could raise the global average sea-level by 7 metres over a period of 1,000 years or more. We show here that concentrations of greenhouse gases will probably have reached levels before the year 2100 that are sufficient to raise the temperature past this warming threshold.

At present, about half of the snow falling on Greenland melts and runs off as water, and the remainder is discharged in the form of icebergs. Climate change caused by higher greenhouse-gas concentrations is expected to produce both higher temperatures and greater precipitation, but most studies conclude that the increase in melting will outweigh the increase in snowfall¹. For an annual average warming of more than 2.7 °C, the melting exceeds the snowfall² — a situation in which the ice-sheet must contract, even if ice-berg production is reduced to zero as it retreats from the coast.

For a warming of 3 °C, the ice-sheet loses mass slowly^{3,4} and over millennia might approach a steady state in a smaller inland form. For greater warming, mass is lost faster and the ice-sheet is likely to melt away. The most extreme scenario considered in the third assessment report of the Intergovernmental Panel on Climate Change (IPCC)¹ involves a warming of 8 °C; in this case, most of the ice-sheet disappears over the next 1,000 years³.

The magnitude of predicted global warming depends on the concentration of greenhouse gases and the climate response to these. We have calculated the development of Greenland's temperature using scenarios⁵ in which atmospheric carbon dioxide concentration stabilizes at different levels over the next few centuries (Fig. 1a). To do this, we used methods from the IPCC report^{1,6}, in which a simple climate model is tuned to reproduce the results of a range of atmosphere–ocean general circulation models (AOGCMs); the range represents uncertainties in climate modelling. We take into account the AOGCMs' prediction that future warming in Greenland will be larger than the global average: warming is greater at high northern latitudes because the loss of snow and sea ice creates positive feedbacks, due in particular to the reduced reflection of sunlight.

The 2.7 °C threshold is passed in all but one of the 35 combinations of AOGCM and stabilization level; the warming exceeds 8 °C in many cases and continues to rise after 2350 for the higher concentrations. The threshold was derived assuming uniform warming

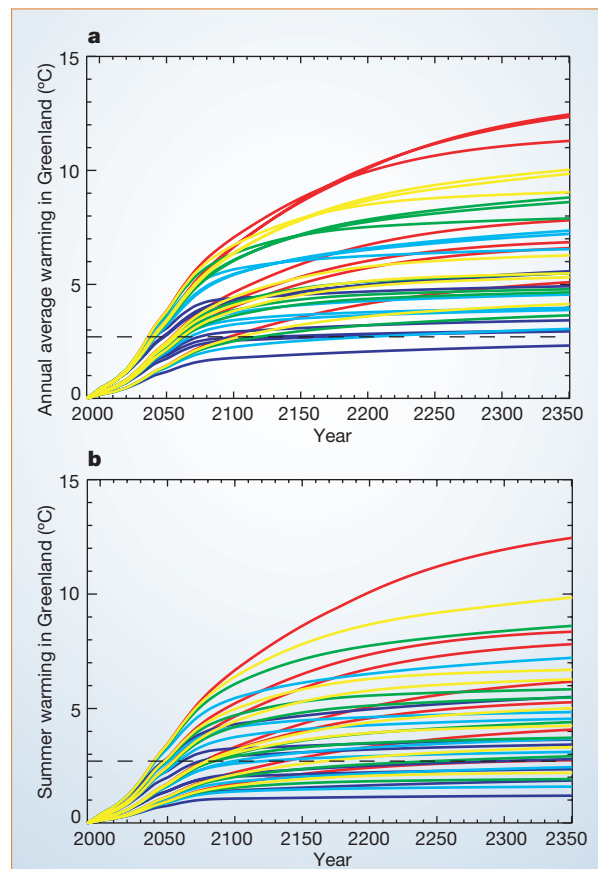


Figure 1 Predicted warming of Greenland over the next few centuries. **a**, **b**, Change over time in **a**, annual average temperature, and **b**, summer temperature in Greenland relative to temperature in 1990, as predicted by seven climate models^{1,6} (shown as seven lines for each colour) for emissions scenarios⁵ in which the atmospheric carbon dioxide concentration stabilizes at five different levels (in p.p.m.: purple, 450; light blue, 550; green, 650; yellow, 750; red, 1,000). For comparison, the pre-industrial concentration of atmospheric carbon dioxide was about 280 p.p.m. and that of the present day stands at about 370 p.p.m. Scenarios involving higher carbon dioxide concentrations stabilize later. Stabilization with a concentration of 1,000 p.p.m. is attained in the year 2375. Dashed lines show the threshold for the viability of the ice-sheet at 2.7 °C warming. This threshold has ± 0.5 °C uncertainty due to estimates (described in the Intergovernmental Panel on Climate Change's third annual report¹) of uncertainties in the change in precipitation over Greenland, the calculation of melting, the geographical distribution of warming and the effects of ice-sheet dynamics.

through the year, but most of the AOGCMs predict more warming in Greenland in winter than in summer. The ice-sheet is sensitive to warming in summer and early autumn, but not in winter, when no melting takes place. Calculations using summer warming instead of annual average warming and allowing for uncertainty in the threshold (Fig. 1b) could reduce the number of cases passing the threshold to 24 out of 35 (69%).

The lowest carbon dioxide concentration considered was 450 p.p.m. Given that this level is exceeded before 2050 in all of the IPCC report's emission scenarios, and that carbon dioxide is not the only greenhouse gas, we conclude that the Greenland ice-sheet is likely to be eliminated by anthropogenic climate change unless much more substantial emission reductions are made than those envisaged by the IPCC. This would mean a global average sea-level rise of 7 metres during the next 1,000 years or more.

Without the ice-sheet, the climate of Greenland would be much warmer because the land surface would be at a lower altitude and reflect less sunlight. This conclusion can be drawn without detailed modelling. Even if atmospheric composition and the global climate were to return to pre-industrial conditions, the ice-sheet might not be regenerated^{7,8}, which implies that the sea-level rise could be irreversible.

Jonathan M. Gregory*†, Philippe Huybrechts‡§, Sarah C. B. Raper§

*Centre for Global Atmospheric Modelling,

Department of Meteorology, University of Reading, Reading RG6 6BB, UK

e-mail: j.m.gregory@reading.ac.uk

†Hadley Centre for Climate Prediction and Research, Met Office, Exeter EX1 3PB, UK

‡Department of Geography, Vrije Universiteit Brussel, 1050 Brussels, Belgium

§Alfred-Wegener-Institut für Polar- und Meeresforschung, 27515 Bremerhaven, Germany

- Church, J. A. et al. in *Climate Change 2001: The Scientific Basis* Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change (eds Houghton, J. T. et al.) 639–693 (Cambridge Univ. Press, 2001).
- Huybrechts, P., Letreguilly, A. & Reeh, N. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* (Global and Planetary Change Section) **89**, 399–412 (1991).
- Huybrechts, P. & De Wolde, J. *J. Clim.* **12**, 2169–2188 (1999).
- Greve, R. *Clim. Change* **46**, 283–289 (2000).
- Wigley, T. M. L. et al. *Nature* **379**, 240–243 (1996).
- Cubasch, U. et al. in *Climate Change 2001: The Scientific Basis* Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change (eds Houghton, J. T. et al.) 525–582 (Cambridge Univ. Press, 2001).
- Crowley, T. J. & Baum, S. K. *Paleoceanography* **10**, 357–363 (1995).
- Toniazzo, T., Gregory, J. M. & Huybrechts, P. *J. Clim.* **17**, 21–33 (2004).

Competing financial interests: declared none.

brieF communications arising online

www.nature.com/bca

Evolutionary biology: Sex change and relative body size in animals

P. M. Biston, P. L. Munday & R. R. Warner (doi:10.1038/nature02512)

Reply: D. J. Allsop & S. A. West (doi:10.1038/nature02513)

Evolutionary biology

Sex change and relative body size in animals

Arising from: Allsop, D. J. & West, S. A. *Nature* **425**, 783–784 (2003)

Organisms that change sex during their lifetime use a variety of strategies — they may be female first¹, male first² or even repetitive sex changers³. Natural selection should favour those individuals that change sex at a time when it increases their reproductive value^{4–6}. Allsop and West⁷ claim that the relative timing of sex change is invariant across all animals, with individuals changing sex at 72% of their maximum body size, and infer that natural selection for sex change must therefore be fundamentally similar across animals. Here we explain why we believe that Allsop and West's claims are not supported by their analysis or by their empirical data⁷.

Inspection of the data underlying Allsop and West's results^{7,8} reveals that relative size at sex change ($L_{50}/L_{\max}$) is highly variable (Fig. 1a). The basis for their claim of invariance is a tight relationship and a slope of unity when the average size at sex change (L_{50}) is plotted against maximum size ($L_{\max}$)⁷. We suggest that the same relationship would hold if the average size at sex change were randomly distributed between $L_{\max}$

the size at maturity, and maximum size.

To test this idea, we developed a null model. Species were randomly assigned a maximum size (between 2 mm and 1.5 m)⁷, a size at maturity (assumed for simplicity to be 50% of their maximum size)⁹, and an average size at sex change between maturation and maximum size. We used this null model to generate 10 data sets with 77 species in each (Fig. 1b). This null model excludes only the factor of interest (the real distribution of size at sex change) while incorporating other realistic factors (such as a non-zero size at maturity) that might confound the results¹⁰.

Our regressions of $\log(L_{50})$ against $\log(L_{\max})$ were indistinguishable from the regression found by Allsop and West⁷. Our analyses gave significant slopes ranging between 0.96 and 1.04, and explained 95–97% of the variation in size at sex change (Fig. 1c). Our results satisfy the criteria used to claim that the relative size at sex change is invariant⁷, even though the relative size at sex change is randomly distributed.

We repeated our analyses with a more realistic null model¹⁰ in which each species was

randomly assigned a size at maturity of between 40% and 80% of its maximum size¹¹, and found that this did not alter our conclusions. In this model, relative size at sex change develops a left skew like that seen in Fig. 1a. Furthermore, the variance in $L_{50}/L_{\max}$ (mean $\pm$ s.d. = 0.018 ± 0.003 ; $n = 10$ data sets) is indistinguishable from the variance found in the real data. The results of the null model of Gardner *et al.* (A. Gardner, E. Charnov, D. J. Allsop and S. A. West, manuscript in preparation) depend on the questionable assumption that size at maturity can range over 0–100% of maximum size. We conclude that Allsop and West's results⁷ do not demonstrate that relative size at sex change is invariant, and therefore that they offer no insight into natural selection for sex change.

These problems arise because size at maturity and maximum size constrain the set of possible values that average size at sex change may take (Fig. 1d). When the relationship between $L_{\max}$ and L_{50} is plotted on a log–log scale, these constraints cause apparent invariance in L_{50} and a restriction in the range of possible slopes to values near 1.0 (Fig. 1e). As constraints on the attribute of interest become more stringent, it will generally become harder to reject the null hypothesis that the attribute is randomly distributed between the constraints.

Empirical data demonstrate that individuals change sex over a large range of sizes^{1–3}. The timing of and size at sex change are often precisely linked to changes in relative condition and group membership^{1–3}, suggesting that natural selection has shaped flexible sex-change strategies that are contingent on social context. To advance our understanding further we need to attend to the great variation in sex-change strategies within and among species.

Peter M. Buston*, Philip L. Munday†, Robert R. Warner‡

*National Center for Ecological Analysis and Synthesis, University of California, Santa Barbara, California 93101, USA

e-mail: buston@nceas.ucsb.edu

†School of Marine Biology and Aquaculture, James Cook University, Townsville, Queensland 4811, Australia

‡Department of Ecology, Evolution, and Marine Biology, University of California, Santa Barbara, California 93106, USA

doi:10.1038/nature02512

- Warner, R. R. & Robertson, D. R. *Smithson. Contr. Zool.* **254**, 1–27 (1978).
- Buston, P. M. *Nature* **424**, 145–146 (2003).
- Munday, P. L., Caley, M. J. & Jones, G. P. *Behav. Ecol. Sociobiol.* **43**, 371–377 (1998).
- Ghiselin, M. T. *Q. Rev. Biol.* **44**, 189–208 (1969).
- Warner, R. R. *Am. Nat.* **109**, 61–82 (1975).
- Charnov, E. L. *The Theory of Sex Allocation* (Princeton University Press, Princeton, 1982).
- Allsop, D. J. & West, S. A. *Nature* **425**, 783–784 (2003).
- Allsop, D. J. & West, S. A. *J. Evol. Biol.* **16**, 921–929 (2003).
- Charnov, E. L. *Life History Invariants* (Oxford University Press, Oxford, 1993).
- Gotelli, N. J. & Graves, G. R. *Null Models in Ecology* (Smithsonian Institution Press, Washington DC, 1996).
- Beverton, R. J. H. *J. Fish Biol.* **B41**, 137–160 (1992).

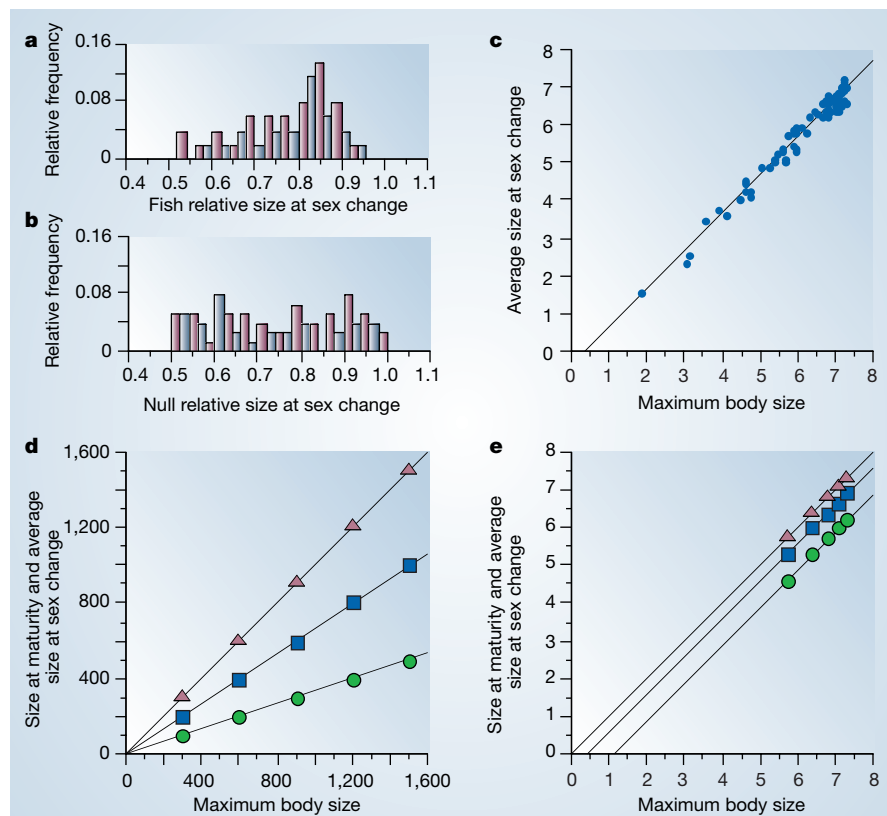


Figure 1 Relative size at sex change varies widely among animals. **a**, Distribution of relative size at sex change ($L_{50}/L_{\max}$) for 52 species of fish used by Allsop and West^{7,8}. **b**, Distribution of relative size at sex change for 77 hypothetical species generated by our null model. **c**, Log–log plot of average size at sex change (L_{50}) against maximum size ($L_{\max}$) for 77 hypothetical species. The null data generate apparent invariance in relative size at sex change (**b**, **c**). **d**, Size at maturity ($L_{\max}$) plotted against maximum size: green circles, $L_{\max} = 33\%$ $L_{\max}$; blue squares, $L_{\max} = 66\%$ $L_{\max}$; red triangles, $L_{\max} = 100\%$ $L_{\max}$. Average size at sex change must fall within these constraints. **e**, Log–log plot of the data shown in **d**. From **d**, **e**, it is evident that more stringent constraints generate more apparent invariance.

Allsop & West reply — Buston *et al.*¹ criticize our² use of standard methodology^{3–5} to test for an invariant relative size at sex change, and propose instead a null model based on randomization techniques: however, their *ad hoc* model is not null.

The main problem is that it assumes an invariant relative size at maturity, which follows from two of the dimensionless invariants assumed by Charnov's model⁶: $\alpha \cdot M$ and k/M (where α is the age at first breeding; M is the adult mortality rate; and k is the relative growth rate, or Bertalanffy coefficient). If these are invariant, then their product $\alpha \cdot k$ is invariant, and so the relative size at maturity ($L_{\text{mat}}/L_{\text{max}} = 1 - \exp(-\alpha \cdot k)$) is also invariant. These are the crucial invariants for Charnov's model, so we would expect the null model of Buston *et al.* to produce an invariant relative size at sex change, and hence to fit our data. If an invariant relative size at maturity is not assumed, then more appropriate null models can be developed (for example, $L_{\text{mat}}/L_{\text{max}} \sim U[0,1]$, $L_{50}/L_{\text{max}} \sim U[L_{\text{mat}}/L_{\text{max}}, 1]$) and the predictions of these differ significantly from the observed data (A. Gardner, E. Charnov, D.J.A. and S.A.W., manuscript in preparation; simulation results, $P < 0.0001$).

There are other problems with the model of Buston *et al.* First, the distribution of relative size at sex change in the actual data is significantly different from the uniform distribution they assume (for fish: $P = 0.000001$; for all species: $P = 0.02$). More generally, invariance is statistical — it does not imply that all individuals do exactly the same thing^{3,4}.

Second, they arbitrarily assign a size at maturity of 50% of maximum body size. This forces their model to fit the data, giving an average size at sex change of 75% of maximum body size (observed is 72%). Their citation⁴ actually suggests that 50% is a lower

bound, with an average value of 65%, which would give a mean size at sex change of 83%, far from the observed.

Third, the assumption of a uniform distribution in relative size at sex change assumes no selection on size at sex change, which is not the case^{4,7,8}. Fourth, the model of Buston *et al.* and our version are both unrealistic 'straw men', easily knocked down, as shown here.

A more powerful and informative approach is to carry out a sensitivity analysis of Charnov's model⁶ and test how variation in different parameters influences the relative size at sex change and its variation. We are doing this (A. Gardner, E. Charnov, D.J.A. and S.A.W., manuscript in preparation) and have found that the prediction of an invariant size at sex change relies primarily on invariance in $\alpha \cdot M$ and k/M , with variation in δ — the coefficient relating male fertility to size — having little effect. This explains the results of the null model of Buston *et al.* (which does not assume an invariant δ) and explains why breeding system and taxa do not significantly influence the relative size at sex change².

Buston *et al.* suggest that variation within species provides problems for our invariant result. However, this criticism misses the purpose of cross-species comparative studies^{5,9,10}. The aim is to look for general patterns across species — this does not imply that there is no within-species variation. For example, α is often facultatively adjusted within species¹¹, but this does not disrupt the $\alpha \cdot M$ invariant⁴. The study of variation across and within species should be seen as complementary approaches, not alternatives. Our results² indicate that Charnov's model⁶ efficiently encapsulates the crucial aspects of the underlying biology.

We agree that relative size and social con-

text will influence the advantage and timing of sex change for specific individuals, and that the importance of this varies across species (in contrast to the assumptions made by Buston *et al.* in their model)^{7,8}. However, our results indicate that these effects may average out, so the relationship between size and fitness can be approximated extremely well by a single positive relationship for each sex. It is a statistical fact that biological details matter for the timing of when individuals change sex, but not for explaining the average pattern across species.

Our approach is based on fundamental assumptions and evolutionary theory developed over the past 30 years^{4,7–9}. The novelty lies in the predictions of these models being phrased in terms of dimensionless qualities that are invariant and determining the consequences for general patterns of sex change^{2,4,6}.

David J. Allsop, Stuart A. West

Institute of Cell, Animal and Population Biology,
University of Edinburgh, Edinburgh EH9 3JT, UK
e-mail: david.allsop@ed.ac.uk

doi:10.1038/nature02513

1. Buston, P. M., Munday, P. L. & Warner, R. R. *Nature* doi:10.1038/nature02512 (2004).
2. Allsop, D. J. & West, S. A. *Nature* **425**, 783–784 (2003).
3. Brown, J. H., West, G. B. & Enquist, B. J. in *Scaling in Biology* (eds Brown, J. H. & West, G. B.) 1–24 (Oxford University Press, Oxford, 2000).
4. Charnov, E. L. *Life History Invariants* (Oxford University Press, Oxford, 1993).
5. Harvey, P. H. & Pagel, M. D. *The Comparative Method in Evolutionary Biology* (eds May, R. M. & Harvey, P. H.) (Oxford University Press, Oxford, 1991).
6. Charnov, E. L. & Skuladottir, U. *Evol. Ecol. Res.* **2**, 1067–1071 (2000).
7. Charnov, E. L. *The Theory of Sex Allocation* (Princeton University Press, Princeton, 1982).
8. Leigh, E. G., Charnov, E. L. & Warner, R. R. *Proc. Natl Acad. Sci.* **73**, 3655–3660 (1976).
9. Frank, S. A. *Foundations of Social Evolution* (Princeton University Press, Princeton, 1998).
10. Allsop, D. J. & West, S. A. *Evolution* (in the press).
11. Stearns, S. C. *The Evolution of Life Histories* (Oxford University Press, Oxford, 1992).

Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*

Manolis Kellis^{1,2}, Bruce W. Birren¹ & Eric S. Lander^{1,3}

¹The Broad Institute, Massachusetts Institute of Technology and Harvard University, Cambridge, Massachusetts 02138, USA

²MIT Computer Science and Artificial Intelligence Laboratory, and ³Whitehead Institute for Biomedical Research, Cambridge, Massachusetts 02139, USA

Whole-genome duplication followed by massive gene loss and specialization has long been postulated as a powerful mechanism of evolutionary innovation. Recently, it has become possible to test this notion by searching complete genome sequence for signs of ancient duplication. Here, we show that the yeast *Saccharomyces cerevisiae* arose from ancient whole-genome duplication, by sequencing and analysing *Kluyveromyces waltii*, a related yeast species that diverged before the duplication. The two genomes are related by a 1:2 mapping, with each region of *K. waltii* corresponding to two regions of *S. cerevisiae*, as expected for whole-genome duplication. This resolves the long-standing controversy on the ancestry of the yeast genome, and makes it possible to study the fate of duplicated genes directly. Strikingly, 95% of cases of accelerated evolution involve only one member of a gene pair, providing strong support for a specific model of evolution, and allowing us to distinguish ancestral and derived functions.

Genomic duplication has been proposed as an advantageous path to evolutionary innovation¹, because duplicated genes can supply genetic raw material for the emergence of new functions through the forces of mutation and natural selection. Such duplication can involve individual genes, genomic segments or whole genomes. Whole-genome duplication (WGD) is a particularly intriguing but poorly understood situation. In principle, coordinate duplication of an entire genome may allow for large-scale adaptation to new environments. However, polyploidy comes at the cost of major genomic instability², which persists until the genome returns to functionally normal ploidy through mutation, gene loss and genomic rearrangements. Direct study of such a cataclysmic genomic event may provide major insights into the dynamics of genome evolution and the emergence of new functions.

In the yeast *S. cerevisiae*, subtle analysis of the genomic locations of paralogues revealed the existence of ancestral duplication blocks, but their origin has remained controversial^{3–17}. Wolfe and colleagues interpreted the presence and distribution of such regions in the *S. cerevisiae* genome as supporting a model of WGD^{3,4}. Others have challenged this conclusion, noting that the WGD model is based on a small set of yeast genes (~8%), explains only a minority of the gene redundancy in the genome and defines putative sister regions covering less than 50% of the genome. An alternative model that has been proposed is that the duplicated segments arose via independent local duplication events^{8–15}. Recent reports assert that low-coverage sequence data from related yeast species supports this alternative model^{14,15}; Wolfe and colleagues have in turn challenged this analysis¹⁶. In addition, Wolfe and colleagues postulated that WGD occurred after the divergence of *S. cerevisiae* and *K. lactis*, while Langkjaer *et al.* have recently presented evidence that any putative WGD would necessarily predate this divergence¹⁷.

Here we provide direct evidence of WGD in yeast, by sequencing and analysing a related species whose divergence precedes the duplication event. We show that *S. cerevisiae* arose from complete duplication of eight ancestral chromosomes, and subsequently returned to functionally normal ploidy by massive loss of nearly 90% of duplicated genes in small deletions. These were balanced and complementary in paired regions, preserving at least one copy of virtually each gene in the ancestral gene set. We identify 145 paired

regions in *S. cerevisiae*, tiling 88% of the genome and containing 457 duplicated gene pairs.

We then analyse the post-duplication divergence of gene pairs, and show evidence of accelerated evolution in many cases. Strikingly, 95% of cases of accelerated evolution involve only one member of a gene pair, providing strong support for a specific model of evolution¹, and allowing ancestral and derived functions to be distinguished. We find that derived genes tend to be specialized in function, expression and localization, and lose essential aspects of their ancestral function. In addition, we find striking examples of neofunctionalization, including the emergence of silencing from origin-of-replication binding, and the emergence of viral defence mechanisms from translation elongation.

Evidence of whole-genome duplication

The expected signature of genome duplication is illustrated in Fig. 1. Following duplication, sister regions would undergo progressive gene loss by random local deletion: one or the other of the two paralogous copies of each gene would be lost in most cases, with both paralogues being retained if they happen to acquire distinct functions. Eventually, the only residual signature to show that two regions arose from ancestral duplication is the presence of a few paralogous genes in the same order and orientation scattered in a sea of otherwise unrelated genes. Paired regions containing such signatures will be relatively short, because chromosomal rearrangements would disrupt global gene order over time, leaving only weak evidence of ancestral duplication.

The clearest way to prove the existence of an ancient WGD would be to find a yeast species, denoted as Y, that descends directly from a common ancestor along a lineage that diverged before the duplication (Fig. 1). In principle, species Y and *S. cerevisiae* would be related by a 1:2 mapping satisfying several properties: (1) nearly every region in species Y would correspond to two sister regions in *S. cerevisiae*; (2) the two sister regions in *S. cerevisiae* would each contain an ordered subsequence of the genes in the corresponding region of species Y, with each containing roughly half of the genes and the two subsequences interleaving to account for nearly all of the genes; and (3) nearly every region of *S. cerevisiae* would correspond to one region of species Y, and thus be paired to a sister

region in *S. cerevisiae*. Below, we show that *K. waltii* matches these criteria.

Genome sequencing and alignment

We sequenced the genome of *K. waltii*, as part of a larger fungal genome programme. We note that *Kluyveromyces* is a polyphyletic genus, and in particular, *K. waltii* is closer to *S. cerevisiae* than the previously studied *K. lactis*. In light of this, *K. waltii* has been proposed to be renamed *Lachancea waltii*¹⁸, although we use the more familiar name here. We used whole-genome shotgun sequencing to obtain eightfold sequence coverage and 42-fold physical coverage, and assembled the information into essentially eight complete chromosomes of total length 10.7 Mb (see Methods).

We annotated the draft genome sequence to identify 5,230 likely protein-coding genes (see Methods), notably fewer than the 5,714 genes in *S. cerevisiae*¹⁹. The genome also contains 240 transfer RNA genes (36 fewer than *Saccharomyces* species) and 60 Ty elements (primarily TyB). It also contains several telomeric gene families that are probably involved in adaptation to changing environments, including families of membrane proteins (51 members), hexose transporters (20 members), and flocculins (50 members, a notable expansion relative to *S. cerevisiae*). Finally, 7% of genes show no detectable protein similarity to *S. cerevisiae*. The genome sequence has been deposited in public databases (GenBank/DBJ/EMBL), and the sequence and annotation are available from our website (<http://www.broad.mit.edu/seq/YeastDuplication>) and as Supplementary Information.

We compared the genomes of *K. waltii* and *S. cerevisiae* to identify orthologous regions. We first identified closely matching genes between the species and then identified blocks of conserved synteny—that is, regions containing numerous matching genes in the same order across the two species (see Methods). In contrast to the 1:1 mapping seen for close relatives¹⁹, most local regions in *K. waltii* mapped to two regions in *S. cerevisiae*, with each containing matches to only a subset of the *K. waltii* genes.

To quantify this observation, we developed a computational algorithm to parse *K. waltii* into blocks of doubly conserved synteny (DCS)—defined as maximal regions that map across their entire length to two distinct regions in *S. cerevisiae* (see Methods). Briefly, we labelled the genes within each synteny block, and used the genes that belonged to multiple synteny blocks to detect regions of duplicate synteny. We then used the duplicate mapping in DCS blocks to define sister regions in the *S. cerevisiae* genome (see Methods).

Complete duplication

We identified a total of 253 DCS blocks, containing 75% of *K. waltii* genes and 81% of *S. cerevisiae* genes. The cross-species mapping of DCS blocks is shown in Fig. 2a, with a detailed view of a region of chromosome 1. DCS blocks tile 85% of each *K. waltii* chromosome in the pattern expected for WGD. Detailed views of all regions are shown in Supplementary Figs S5.

The typical gene falls in a DCS block containing 27 genes, and the

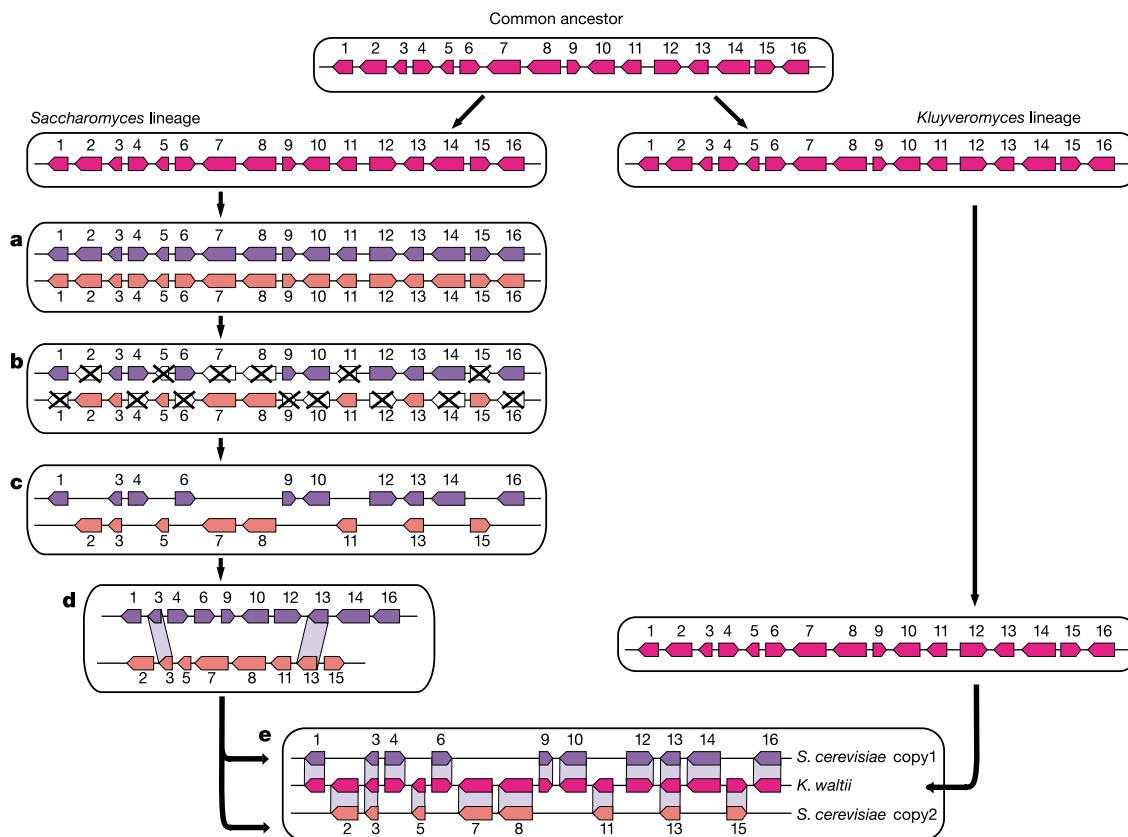


Figure 1 Model of WGD followed by massive gene loss predicts gene interleaving in sister regions. **a**, After divergence from *K. waltii*, the *Saccharomyces* lineage underwent a genome duplication event, creating two copies of every gene and chromosome. **b**, The vast majority of duplicated genes underwent mutation and gene loss. **c**, Sister segments retained different subsets of the original gene set, keeping two copies for only a small minority of duplicated genes, which were retained for functional purposes. **d**, Within

S. cerevisiae, the only evidence comes from the conserved order of duplicated genes (numbered 3 and 13) across different chromosomal segments; the intervening genes are unrelated. **e**, Comparison with *K. waltii* reveals the duplicated nature of the *S. cerevisiae* genome, interleaving genes from sister segments on the basis of the ancestral gene order.

largest such block contains 81 genes. DCS blocks are separated by small segments (~3 genes on average) that appear to match only one conserved region in *S. cerevisiae*; it is likely that the second paralogous segment cannot be confidently recognized because it has lost too many genes or undergone rearrangement (see Methods). Finally, ~1% of the *K. waltii* genome lies in segments that match

three or more regions in *S. cerevisiae*; the additional matching regions result mostly from slight local rearrangements (see Methods).

Complete interleaving of genes

Within DCS blocks, the interleaving of genes shows essentially

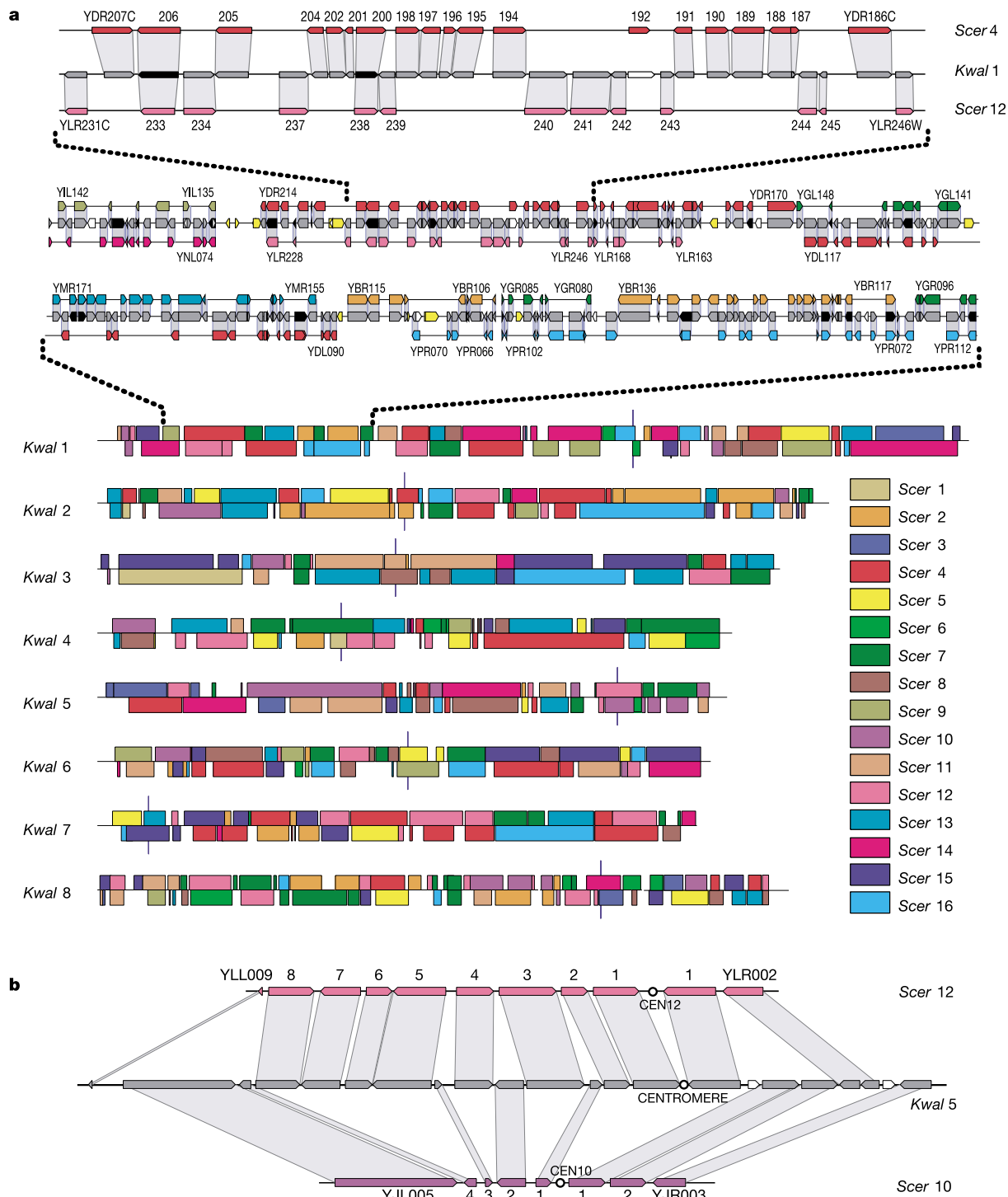


Figure 2 Gene and region correspondence with *K. waltii* reveals WGD. **a**, Each region of *K. waltii* (coloured by number of matches: white, 0; grey, 1; black, 2; yellow, ≥ 3) shows conserved gene order with two regions of *S. cerevisiae* (coloured by chromosome number). Spacing between *S. cerevisiae* genes is set to match *K. waltii* chromosomal positions. Vertical blue bars denote centromeres. **b**, DCS region showing duplicate mapping of centromeres (black circles). All sixteen *S. cerevisiae* centromeres show such

duplicate mapping with *K. waltii* centromeres. This DCS region also illustrates that we can reliably recognize anciently duplicated segments even in the absence of any remaining two-copy genes. Evidence of WGD comes from gene interleaving and 2:1 mapping with orthologous *K. waltii* segments. The segments containing intervening genes are deleted, resulting in condensed sister regions. *Kwal*, *K. waltii*; *Scer*, *S. cerevisiae*.

complete duplicate mapping (Fig. 2b and Supplementary Information). In a typical DCS block, 90% of *K. waltii* genes have matches in at least one of the two *S. cerevisiae* regions (and on average ~12% have matches in both regions). Similarly, in each DCS block, 90% of genes from both *S. cerevisiae* regions have matches within the single corresponding region of *K. waltii*, and they are interleaved onto *K. waltii* while preserving gene order and orientation. The mapping fits the pattern expected for WGD followed by massive gene loss.

Complete pairing of *S. cerevisiae*

Based on the cross-species correspondence, we identified sister regions within *S. cerevisiae* (Fig. 3). We used the DCS blocks to define 253 initial sister regions for which the ancestral gene order

was preserved in all three matching segments (*K. waltii* and the two sister regions in *S. cerevisiae*). We then merged consecutive sister regions when both copies in *S. cerevisiae* preserved the ancestral gene order, but were apparently separated owing to a rearrangement in *K. waltii*. This resulted in 145 sister regions in *S. cerevisiae* covering 88% of the genome.

Many of the paired regions could not be recognized without the relationship to *K. waltii*. For example, 47 DCS regions contain no duplicated genes whatsoever and the evidence of duplication comes solely from the interleaving of genes (Fig. 2b).

The chromosome structure of the two genomes provides further evidence of WGD. Indeed, the sixteen centromeres of *S. cerevisiae* are paired in a 2:1 manner with the eight centromeres of *K. waltii* (Fig. 2b and Methods).

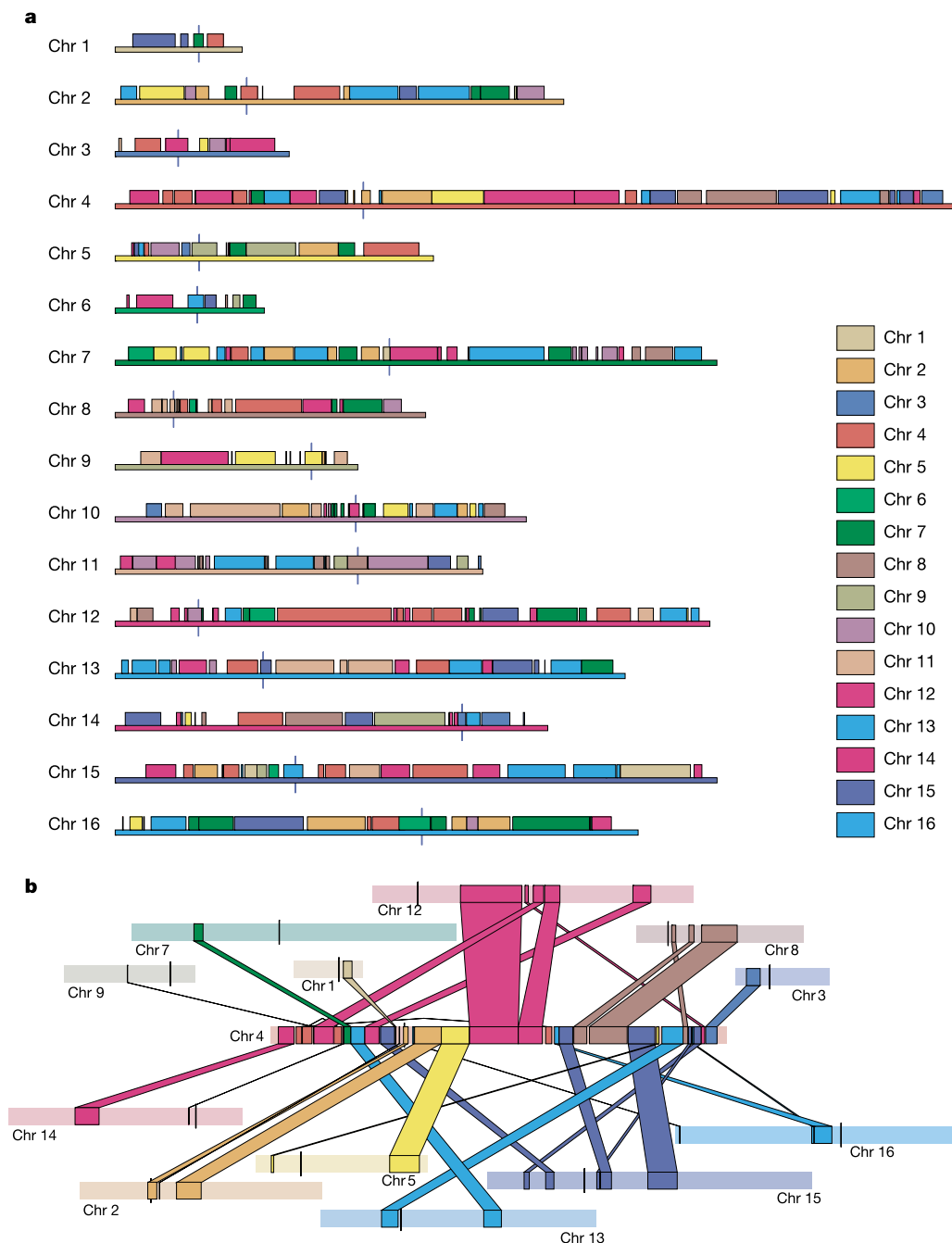


Figure 3 Duplicated blocks in *S. cerevisiae*. **a**, The duplicate mapping in sister regions extends to tile each *S. cerevisiae* chromosome, revealing complete duplication. Blue

vertical bars denote centromeres. **b**, Detailed mapping of chromosome 4 with sister regions in other chromosomes.

Summary

We conclude that a WGD event occurred in the *Saccharomyces* lineage after the divergence from *K. waltii* and that *S. cerevisiae* is therefore a degenerate tetraploid. WGD may have occurred either at the haploid or the diploid stage and either by endo-duplication (auto-polyploidy) or fusion of two close relatives (allo-polyploidy).

Evolutionary analysis

The relationship between *K. waltii* and *S. cerevisiae* provides the first comparison across an ancient WGD event and offers the opportunity to study the long-term fate of a genome after duplication.

Pattern of gene loss

The WGD event doubled the number of chromosomes in the *Saccharomyces* lineage, but subsequent gene-loss events led to the current *S. cerevisiae* genome, which is only 13% larger than *K. waltii* and contains only 10% more genes. The polyploid genome returned to functional normal ploidy, not by meiosis or chromosomal loss, but instead by a large number of deletion events. We can infer that 12% of the paralogous gene pairs were retained in each DCS block, and the remaining 88% of paralogous genes were lost.

In principle, gene loss can occur by large segmental deletions or

individual gene deletions, and can be balanced between the two sisters or act primarily on one of them. Analysis of DCS blocks revealed that gene loss occurred by many small deletions (the average size of a lost segment is two genes), and was typically balanced between the two sister regions (average balance 57% to 43%). Additionally, a large number of rearrangement events have shuffled the duplicated chromosomes; the intergenic regions at the breakpoints seem to be associated with tRNA genes.

Accelerated protein divergence

Most interestingly, we can study the evolution of the 457 gene pairs that arose by WGD. The synteny information allows us to unambiguously establish the ancestry for these gene pairs and distinguish them from many gene pairs in the yeast genome that arose by local duplication events (see Methods). This allows us to study their evolution with respect to each other and the non-duplicated *K. waltii* orthologue.

Two alternatives have been proposed for post-duplication divergence of duplicated gene pairs. Ohno has hypothesized that after duplication, one copy would preserve the original function and the other copy would be free to diverge¹. However, this model has been contested by others, who argued that both copies would diverge more rapidly and acquire new functions^{20–22}. We calculated divergence rates for the 457 pairs, using sequence information from *K. waltii*, *S. cerevisiae* and the related yeast *S. bayanus* (Fig. 4d). We searched for instances of unusually rapid or slow evolution.

We found that 76 of the 457 gene pairs (17%) show accelerated protein evolution relative to *K. waltii*, defined as instances in which the amino acid substitution rate along one or both of the *S. cerevisiae* branches was at least 50% faster than the rate along the *K. waltii* branch (Tables 1–3). We note that this calculation is conservative and may miss some cases of accelerated evolution (see Methods). These genes were heavily biased towards protein kinases (13 pairs, P -value 10^{-8}) and regulatory proteins (8 pairs, including the cell-cycle transcription factors Swi5 and Ace2, and the filamentation factors Phd1 and Sok2), and were generally involved in metabolism (38 pairs, P -value 10^{-18}) and cell growth (32 pairs, P -value 10^{-10}).

Ancestral and derived functions

Strikingly, in nearly every case (95%), accelerated evolution was confined to only one of the two paralogues. This strongly supports the model in which one of the paralogues retained an ancestral function while the other, relieved of this selective constraint, was free to evolve more rapidly¹.

For these 76 gene pairs showing accelerated evolution with respect to *K. waltii*, we can infer that the slowly evolving paralogue has probably retained the ancestral gene function and that the rapidly evolving paralogue probably acquired a derived function after duplication. In fact, the inference of ancestral and derived status can be extended to a total of 115 gene pairs consisting of those for which one *S. cerevisiae* paralogue has evolved at least 50% faster than the other *S. cerevisiae* paralogue.

This analysis thus identifies newer functions that evolved from older ones. We can infer, for example, that the silencing function of Sir3 (which acts in telomeres and mating type cassettes)²³ derived from the origin-of-replication binding function of Orc1 (ref. 24). Similarly, the anti-viral function of Ski7 that recognizes and represses non-polyadenylated messenger RNA was probably derived from the translation-elongation function of Hbs1 (ref. 25).

In many cases, the derived paralogues are specialized in their cellular localization or temporal expression. Acc1 (acetyl-CoA carboxylase) gave rise to Hfa1, which now functions in the mitochondrion²⁶. Skt5, involved in cell-wall biosynthesis, gave rise to Shc1, whose function is restricted to spore formation²⁷. And the pleiotropic transcription factor Skn7, involved in a number of stress responses, gave rise to the transcription factor Hms2, involved in pseudohyphal growth under nitrogen depletion²⁸.

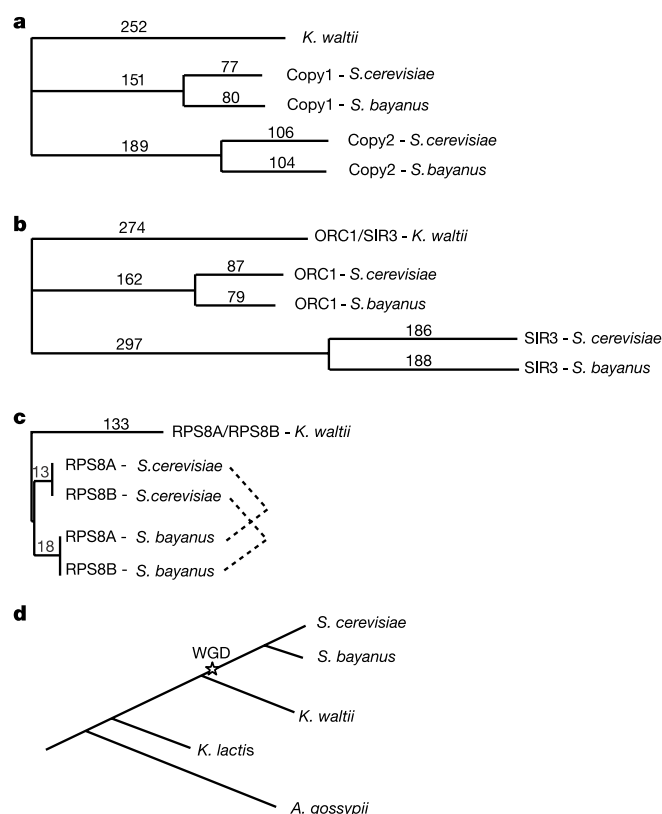


Figure 4 Divergence of duplicated gene pairs. Branches show number of substitutions per thousand amino acids. Evolutionary trees are rooted at the divergence of three species. **a**, Average protein divergence of all 457 gene pairs that arose by WGD. The faster-evolving paralogue is arbitrarily designated as copy 2 for each pair. **b**, Example of a gene showing accelerated protein divergence (1 of 72 cases). Ancestral and derived gene function can be inferred by comparison to *K. waltii*. In this case, the origin-of-replication recognition complex protein Orc1 is inferred to be ancestral and the silencing protein Sir3 is inferred to be derived. **c**, Example of duplicated gene pairs that have undergone recent gene conversion (1 of 60 cases). Comparison with *S. bayanus* shows that recent gene conversion events occurred both in *S. cerevisiae* and in *S. bayanus* lineages. Dotted lines connect orthologous genes in *S. cerevisiae* and in *S. bayanus*. **d**, Phylogeny and relative time of WGD. Estimated tree lengths are as reported in ref. 45.

Table 1 **Rapid protein divergence**

Slowly evolving gene (ancestral function)		Rapidly evolving gene (derived function)		Kwal rate	Relative rates	
Gene 1	Description 1	Gene 2	Description 2		Rate1 (%)	Rate2 (%)
<i>RNR2</i>	Small subunit of ribonucleotide reductase	<i>RNR4</i>	Ribonucleotide reductase	0.069	146	519
<i>MCK1</i>	Serine/threonine/tyrosine protein kinase	<i>YKG3</i>	Yeast homologue of mammalian glycogen synthase kinase 3	0.102	110	479
<i>SEC14</i>	Phosphatidylinositol transfer protein	<i>SFH1</i>	Sec14p homologue with highest sequence similarity but lowest functional redundancy	0.061	149	462
<i>YSH1</i>	Subunit of polyadenylation factor I (PFI)	<i>SYC1</i>	C-terminal portion of YSH1	0.130	112	380
<i>GCS1</i>	ADP-ribosylation factor GTPase-activating protein (ARF GAP)	<i>SPS18</i>	Transcription factor	0.172	112	309
<i>FEN1</i>	Probable subunit of 1,3-beta-glucan synthase	<i>ELO1</i>	Elongation enzyme 1	0.111	106	293
<i>CET1</i>	mRNA capping enzyme beta subunit (80 kDa)	<i>CTL1</i>	RNA triphosphatase, no capping function	0.231	76	263
<i>EGD1</i>	Regulator of pol II-transcribed genes	<i>BTT1</i>	Negative regulator of RNA polymerase II	0.146	109	240
<i>SKN7</i>	Transcription factor involved in oxidative stress	<i>HMS2</i>	Heat-shock transcription factor homologue	0.243	103	229
<i>HBS1</i>	Translation elongation factor	<i>SKI7</i>	Anti-viral protein	0.241	116	223
<i>SEC24</i>	Vesicle coat component	<i>SFB2</i>	Putative zinc finger protein	0.130	114	203
<i>IFH1</i>	Overexpression interferes with silencing at telomeres / HM, Interacts with FHL1	<i>YDR223W</i>	Hypothetical ORF	0.269	91	189
<i>SKT5</i>	Probable Ca ²⁺ binding membrane protein (prenylated)	<i>SHC1</i>	Sporulation-specific homologue of csd4	0.242	93	188
<i>ORC1</i>	Largest subunit of origin recognition complex	<i>SIR3</i>	Silencing regulator HML, HMR, telomeres	0.274	91	177
<i>BUB1</i>	Serine/threonine protein kinase required for arrest in loss of microtubule function	<i>MAD3</i>	Spindle checkpoint complex subunit	0.249	133	172
<i>ECM18</i>	Involved in cell wall biogenesis	<i>ICT1</i>	Increased copper tolerance	0.224	98	161

Accelerated protein divergence of duplicated gene pairs (16 of 76 pairs). Kwal, amino acid substitution rate along *K. waltii* lineage. Rate1 and Rate2, substitution rates for Gene1 and Gene2 as a percentage of Kwal.

The functional distinction between the derived and ancestral paralogue is confirmed by the phenotype of deletion mutants when grown in rich medium conditions. For each of the 115 gene pairs, we examined the reported results of systematic deletion experiments^{29,30}. Strikingly, deletion of the ancestral paralogue was lethal in 18% of cases (similar to the overall average for all *S. cerevisiae* genes), whereas deletion of the derived paralogue was never lethal. The derived paralogue is thus not essential under these conditions, either because it does not function in rich medium or because the ancestral paralogue can complement its function. Moreover, we infer that along with possibly gaining a new function, the derived copy has lost some essential aspect of its function, and cannot typically complement deletion of the ancestral gene.

Changes in codon usage

We also found that 32 of the 457 gene pairs show accelerated nucleotide, but not protein, evolution (Table 2). For example, the transcription elongation genes *YEF3* and *HEF3* encode very similar proteins and show a rate of protein change similar to that of *K. waltii*. However, *HEF3* shows twofold acceleration in degenerate third-codon positions (relative to both *YEF3* and *K. waltii*), an unusual pattern of codon usage (codon adaptation index³¹ is 0.17 as compared to 0.78 for *YEF3*) and a much lower level of expression³². The same phenomenon is seen for the pyruvate kinase genes *CDC19* and *PYK2*. The accelerated divergence appears to reflect a relief from selection in the codon usage of the accelerated copy, and the presence of both gene copies may allow the cell to sense changes in metabolic state. Interestingly, *PYK2* is expressed preferentially in conditions of low glycolytic flux³³.

Gene conversion

At the other end of the spectrum, 60 of the 457 pairs show decelerated protein evolution, defined as instances where one or both copies evolved at least 50% slower than *K. waltii* (Table 3). These cases often involve proteins known to be highly constrained, such as ribosomal proteins (25 pairs), histone proteins (2 pairs) and translation initiation/elongation factors (4 pairs). In the vast majority of cases (90%), the paralogues both show decelerated evolution and tend to be very similar (98% amino acid identity versus 55% for all pairs), suggesting that they may be subject to periodic gene conversion. Indeed, the strong nucleotide similarity in fourfold degenerate codon positions that are not under the same selective pressures for conservation provides evidence supporting gene conversion (>90% nucleotide identity versus 41% for all pairs). Furthermore, in about half of these cases, the two paralogues in *S. cerevisiae* are closer in sequence to each other than either is to its syntenic orthologue in *S. bayanus*, showing that gene conversion has occurred after the relatively recent divergence of the two *Saccharomyces* lineages (Fig. 4c). For strongly conserved gene pairs, the two copies may confer an evolutionary benefit by serving as a repository for gene conversion and thus buffering deleterious changes in either copy.

Pairs with similar rates

The remaining 321 of the 457 gene pairs did not meet our stringent criteria for accelerated or decelerated evolution. In some cases, the functional changes may be similar to those above but subtler. In other cases, the primary functional divergence may be in regulatory sequences. This seems a likely possibility because the extensive gene

Table 2 **Rapid nucleotide divergence**

Gene 1	Gene 2	Description	Codon usage		Protein identity (%)	Rate in fourfold degenerate sites		
			CAI1	CAI2		Rate1	Rate2	Acceleration (%)
<i>CDC19</i>	<i>PYK2</i>	Pyruvate kinase	0.89	0.13	93	0.22	0.63	283
<i>ADH1</i>	<i>ADH5</i>	Alcohol dehydrogenase	0.81	0.25	91	0.35	0.57	164
<i>YEF3</i>	<i>HEF3</i>	Transcription elongation factor	0.78	0.17	92	0.34	0.62	181
<i>GND1</i>	<i>GND2</i>	Phosphogluconate dehydrogenase	0.62	0.20	96	0.39	0.64	166
<i>GDH1</i>	<i>GDH3</i>	NADP-specific glutamate dehydrogenase	0.59	0.16	95	0.41	0.62	150

Accelerated nucleotide divergence of duplicated gene pairs. CAI, codon adaptation index³¹, a measure of typical codon usage between 0 and 1. Protein identity was measured between Gene1 and Gene2. Rate1 and Rate2, nucleotide substitution rate in fourfold degenerate sites for conserved amino acid residues. Acceleration, relative acceleration of Rate2 compared to Rate1.

Table 3 **Slowly evolving genes**

Gene 1	Gene 2	Description	Within <i>S. cerevisiae</i>		To <i>K. waltii</i>	
			All (%)	Fourfold (%)	All (%)	Fourfold (%)
<i>TEF2</i>	<i>TEF1</i>	Translational elongation factor EF-1 alpha	100	100	87	74
<i>TIF2</i>	<i>TIF1</i>	Translation initiation factor eIF4A	100	99	83	68
<i>RPL1B</i>	<i>RPL1A</i>	Ribosomal protein L1B	99	97	88	67
<i>YCL073C</i>	<i>YKR106W</i>	Hypothetical ORF	99	97	51	25
<i>HHF1</i>	<i>HHF2</i>	Histone H4 (HHF1 and HHF2 code for identical proteins)	97	92	89	71

Gene conversion events (5 of 60 cases). Nucleotide identity is shown within *S. cerevisiae* (between Gene1 and Gene2) and to *K. waltii* (between Gene2 and *K. waltii*). All, overall nucleotide divergence. Fourfold, nucleotide divergence in fourfold degenerate sites encoding conserved amino-acid residues. Supplementary Information contains complete divergence information for all gene pairs.

loss following WGD would probably juxtapose new combinations of intergenic and genic regions. In still other cases, the gene pairs may have been retained primarily to increase gene dosage, as with the slowly evolving genes.

Discussion

WGD followed by massive gene loss and gene specialization offers an important path for large-scale evolutionary innovation. Compared to multiple independent duplications and divergence of individual genes or segments, WGD may be more efficient and may offer great opportunities for coordinated evolution. Although organisms clearly can undergo WGD resulting in complete polyploids (as evidenced by existing tetraploid species of plants and animals^{34–37}), it has been unclear whether WGD can then be followed by massive genome reshaping to yield diploids with expanded gene content.

Our analysis of *K. waltii* definitively proves that *S. cerevisiae* is the descendant of an ancient WGD, as originally proposed by Wolfe³ on the basis of subtle genomic patterns. Comparison with *K. waltii* allows rigorous study of the many evolutionary innovations that arose in this dramatic event. We note that a similar analysis of the *Ashbya gossypii* genome has been carried out (P. Philippsen, personal communication). The results here suggest that it may also be fruitful to search for similar genomic signatures of ancient WGD in other organisms^{38–40}. It will be interesting to see just how far such distant echoes of genomic upheaval may be traced. □

Methods

See the Supplementary Information for assembly contigs and scaffolds (Supplementary Table S1), predicted protein-coding genes (Supplementary Information S2), complete gene correspondence (Supplementary Information S3), a table of genes in syntenic segments (Supplementary Tables S4), gene-by-gene mapping for each chromosome (Supplementary Figs S5), detailed visualization of 253 DCS regions (Supplementary Figs S6), 145 paired regions in *S. cerevisiae* by chromosome (Supplementary Figs S7), alignments of duplicated gene pairs (Supplementary Information S8), protein and nucleotide divergence rates (Supplementary Tables S9) and mapping of centromeres (Supplementary Figs S10), also available from <http://www.broad.mit.edu/seq/YeastDuplication>.

Strains, sequencing and assembly

The sequence and annotation for *S. cerevisiae* reference strain S288C was obtained from <http://www.yeastgenome.org>, and updated according to recent comparative work¹⁹. *K. waltii* strain NCYC2644 was provided by I. Roberts (Norwich Research Park, UK). The WGS sequence was obtained at the Whitehead Institute/MIT Center for Genome Research. We used paired-end sequencing with 7.4-fold coverage in 4-kb plasmid clones and 0.6-fold coverage in 40-kb fosmid clones, providing eightfold sequence coverage and 42-fold physical coverage, with laboratory protocols as described at <http://www.broad.mit.edu>. We used the Arachne^{41,42} program to assemble the resulting sequence information into contigs (continuous blocks of uninterrupted sequence) and scaffolds (contigs linked by paired forward-reverse reads from the same clone). The resulting draft genome sequence has a total length of 10.7 megabases, with ~96% contained in ten large scaffolds containing relatively short sequence gaps (average size 144 base pairs, together comprising only 0.5% of total scaffold length). Seven of the scaffolds appear to correspond to complete chromosomes, with recognizable centromeric and telomeric regions, and the remaining three scaffolds appear to belong to a single chromosome, with the order and orientation of the scaffolds readily inferred from the sequence (see below).

Centromeres and telomeres

The conservation of gene order in syntenic regions surrounding the centromeres allowed us to pinpoint candidate centromere-containing intergenic intervals within *K. waltii*.

Sequence analysis of these intervals revealed a centromere structure similar to that of *S. cerevisiae*, containing three elements CDEI–CDEII–CDEIII, where CDEI is TCACCTG/TCANGTG, CDEII is 100–115 nucleotides with AT > 78%, and CDEIII is TTCCGA. This sequence occurs specifically only at the intervals we determined. Each *K. waltii* centromere maps to exactly two *S. cerevisiae* centromeres, as follows: Cen1:Cen7, Cen2:Cen4, Cen3:Cen14, Cen5:Cen9, Cen6:Cen16, Cen8:Cen11, Cen10:Cen12 and Cen13:Cen15. This pairing agrees with previous reports¹⁶. Telomeric repeat elements were identified in the ends of all scaffolds, except where joins are proposed. On the basis of centromere and telomere locations, paired fosmid ends, and synteny with *S. cerevisiae*, we infer that the three scaffolds together form the eighth *K. waltii* chromosome.

Annotation and gene correspondence

We determined the complete set of all predicted open reading frames (pORFs) in *K. waltii* that start with a methionine and have the potential to encode at least 50 amino acids. We found 6,753 non-overlapping ORFs, of which 5,230 are likely to represent functional protein-coding genes. These contain 4,711 ORFs of at least 150 amino acids, of which 92% show protein similarity to *S. cerevisiae*. From the remaining 2,042 ORFs between 50 and 150 amino acids, 519 show protein similarity to *S. cerevisiae*, and we expect these to encode functional proteins. We estimate that only an additional 45 ORFs at this range have been missed (0.8% of the total), on the basis of the percentage of longer ORFs that show similarity. We then constructed a bipartite graph of protein similarities across the two species, and resolved the gene correspondence based on protein similarity and gene order information using the BUS algorithm⁴³.

Doubly conserved synteny regions

We established blocks of conserved synteny by clustering matches of neighbouring genes that are less than 20 kb apart in each genome. We found 353 synteny blocks containing at least three conserved genes each, and matching a total of 4,590 *S. cerevisiae* genes and 4,139 *K. waltii* ORFs. We treated every synteny block as defining an interval on the *K. waltii* genome, and we marked all genes whose coordinates were contained within that interval. We then searched for *K. waltii* genes that were marked as belonging to multiple blocks, thus defining regions of overlap between synteny blocks (DCS regions).

Establishing ancestry of duplicated genes

We determined the ancestry of duplicated gene pairs solely on the basis of gene order information in DCS regions. The difficulty of reliably establishing ancestry based on amino acid divergence, especially in the presence of accelerated evolution, is illustrated by previous reports that have erroneously established WGD as preceding divergence from *K. lactis*¹⁷ (the correct topology is shown in Fig. 4d), or erroneously inferred that GRS2 would represent the ancestral form of the GRS1/GRS2 pair⁴⁴ (in fact, GRS2 is the derived copy, and shows fourfold amino acid acceleration). In both cases, variation in protein divergence rates has misrepresented the true phylogenetic relationships.

Establishing accelerated divergence

Having established gene ancestry based on gene order information, we used amino acid and nucleotide divergence rates to detect cases of accelerated evolution. In seven of the 457 gene pairs, one or both copies in *S. cerevisiae* are split into multiple ORFs and were not used. We used CLUSTALW to construct protein, nucleotide, and codon-aware nucleotide alignments. We rooted three-way trees at the central node, whereas the true root lies on the *K. waltii* branch, and hence our criteria for accelerated divergence are conservative because they overestimate the *K. waltii* branch length. This is especially true in the case of gene-conversion events, and additional pre-duplication relatives will be necessary to estimate divergence between duplication and gene conversion.

Received 17 December 2003; accepted 19 January 2004; doi:10.1038/nature02424.

Published online 7 March 2004.

- Ohno, S. *Evolution by Gene Duplication* (Allen and Unwin, London, 1970).
- Mayer, V. W. & Aguilera, A. High levels of chromosome instability in polyploids of *Saccharomyces cerevisiae*. *Mutat. Res.* **231**, 177–186 (1990).
- Wolfe, K. H. & Shields, D. C. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387**, 708–713 (1997).
- Seoighe, C. & Wolfe, K. H. Updated map of duplicated regions in the yeast genome. *Gene* **238**, 253–261 (1999).
- Melnick, L. & Sherman, F. The gene clusters ARC and COR on chromosomes 5 and 10, respectively, of *Saccharomyces cerevisiae* share a common ancestry. *J. Mol. Biol.* **233**, 372–388 (1993).
- Goffeau, A. *et al.* Life with 6000 genes. *Science* **274**, 546, 563–567 (1996).

7. Philippsen, P. *et al.* The nucleotide sequence of *Saccharomyces cerevisiae* chromosome XIV and its evolutionary implications. *Nature* **387**, 93–98 (1997).
8. Mewes, H. W. *et al.* Overview of the yeast genome. *Nature* **387**, 7–65 (1997).
9. Coissac, E., Maillier, E. & Netter, P. A comparative study of duplications in bacteria and eukaryotes: the importance of telomeres. *Mol. Biol. Evol.* **14**, 1062–1074 (1997).
10. Friedman, R. & Hughes, A. L. Gene duplication and the structure of eukaryotic genomes. *Genome Res.* **11**, 373–381 (2001).
11. Hughes, T. R. *et al.* Widespread aneuploidy revealed by DNA microarray expression profiling. *Nature Genet.* **25**, 333–337 (2000).
12. Piskur, J. Origin of the duplicated regions in the yeast genomes. *Trends Genet.* **17**, 302–303 (2001).
13. Koszul, R., Caburet, S., Dujon, B. & Fischer, G. Eukaryotic genome evolution through the spontaneous duplication of large chromosomal segments. *EMBO J.* **23**, 234–243 (2004).
14. Llorente, B. *et al.* Genomic exploration of the hemiascomycetous yeasts: 18. Comparative analysis of chromosome maps and synteny with *Saccharomyces cerevisiae*. *FEBS Lett.* **487**, 101–112 (2000).
15. Llorente, B. *et al.* Genomic exploration of the hemiascomycetous yeasts: 20. Evolution of gene redundancy compared to *Saccharomyces cerevisiae*. *FEBS Lett.* **487**, 122–133 (2000).
16. Wong, S., Butler, G. & Wolfe, K. H. Gene order evolution and paleopolyploidy in hemiascomycete yeasts. *Proc. Natl Acad. Sci. USA* **99**, 9272–9277 (2002).
17. Langkjaer, R. B., Cliften, P. F., Johnston, M. & Piskur, J. Yeast genome duplication was followed by asynchronous differentiation of duplicated genes. *Nature* **421**, 848–852 (2003).
18. Kurtzman, C. P., Lynch, M. & Force, A. Phylogenetic circumscription of *Saccharomyces*, *Kluyveromyces* and other members of the Saccharomycetaceae, and the proposal of the new genera *Lachancea*, *Nakaseomyces*, *Naumovia*, *Vanderwaltozyma* and *Zygorulasporea*. *FEMS Yeast Res.* **4**, 233–245 (2003).
19. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E. S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**, 241–254 (2003).
20. Lynch, M. & Force, A. The probability of duplicate gene preservation by subfunctionalization. *Genetics* **154**, 459–473 (2000).
21. Force, A. *et al.* Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**, 1531–1545 (1999).
22. Kondrashov, F. A., Rogozin, I. B., Wolf, Y. I. & Koonin, E. V. Selection in the evolution of gene duplications. *Genome Biol.* **3**, RESEARCH0008.1–0008.9 (2002).
23. Shore, D., Squire, M. & Nasmyth, K. A. Characterization of two genes required for the position-effect control of yeast mating-type genes. *EMBO J.* **3**, 2817–2823 (1984).
24. Bell, S. P., Kobayashi, R. & Stillman, B. Yeast origin recognition complex functions in transcription silencing and DNA replication. *Science* **262**, 1844–1849 (1993).
25. Benard, L., Carroll, K., Valle, R. C., Masison, D. C. & Wickner, R. B. The ski7 antiviral protein is an EF1- α homolog that blocks expression of non-Poly(A) mRNA in *Saccharomyces cerevisiae*. *J. Virol.* **73**, 2893–2900 (1999).
26. Huh, W. K. *et al.* Global analysis of protein localization in budding yeast. *Nature* **425**, 686–691 (2003).
27. Sanz, M., Trilla, J. A., Duran, A. & Roncero, C. Control of chitin synthesis through Shc1p, a functional homologue of Chs4p specifically induced during sporulation. *Mol. Microbiol.* **43**, 1183–1195 (2002).
28. Lorenz, M. C. & Heitman, J. Regulators of pseudohyphal differentiation in *Saccharomyces cerevisiae* identified through multicopy suppressor analysis in ammonium permease mutant strains. *Genetics* **150**, 1443–1457 (1998).
29. Winzler, E. A. *et al.* Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285**, 901–906 (1999).
30. Gu, Z. *et al.* Role of duplicate genes in genetic robustness against null mutations. *Nature* **421**, 63–66 (2003).
31. Sharp, P. M. & Li, W. H. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15**, 1281–1295 (1987).
32. Sarthy, A. V. *et al.* Identification and kinetic analysis of a functional homolog of elongation factor 3, YEF3 in *Saccharomyces cerevisiae*. *Yeast* **14**, 239–253 (1998).
33. Boles, E. *et al.* Characterization of a glucose-repressed pyruvate kinase (Pyk2p) in *Saccharomyces cerevisiae* that is catalytically insensitive to fructose-1,6-bisphosphate. *J. Bacteriol.* **179**, 2987–2993 (1997).
34. Hughes, M. K. & Hughes, A. L. Evolution of duplicate genes in a tetraploid animal, *Xenopus laevis*. *Mol. Biol. Evol.* **10**, 1360–1369 (1993).
35. Gallardo, M. H., Bickham, J. W., Honeycutt, R. L., Ojeda, R. A. & Kohler, N. Discovery of tetraploidy in a mammal. *Nature* **401**, 341 (1999).
36. Bailey, G. S., Poulter, R. T. & Stockwell, P. A. Gene duplication in tetraploid fish: model for gene silencing at unlinked duplicated loci. *Proc. Natl Acad. Sci. USA* **75**, 5575–5579 (1978).
37. Otto, S. P. & Whitton, J. Polyploid incidence and evolution. *Annu. Rev. Genet.* **34**, 401–437 (2000).
38. Blanc, G., Barakat, A., Guyot, R., Cooke, R. & Delseny, M. Extensive duplication and reshuffling in the *Arabidopsis* genome. *Plant Cell* **12**, 1093–1101 (2000).
39. Sidow, A. Gen(om)e duplications in the evolution of early vertebrates. *Curr. Opin. Genet. Dev.* **6**, 715–722 (1996).
40. Pebusque, M. J., Coulier, F., Birnbaum, D. & Pontarotti, P. Ancient large-scale genome duplications: phylogenetic and linkage analyses shed light on chordate genome evolution. *Mol. Biol. Evol.* **15**, 1145–1159 (1998).
41. Batzoglou, S. *et al.* ARACHNE: a whole-genome shotgun assembler. *Genome Res.* **12**, 177–189 (2002).
42. Jaffe, D. B. *et al.* Whole-genome sequence assembly for mammalian genomes: arachne 2. *Genome Res.* **13**, 91–96 (2003).
43. Kellis, M., Patterson, N., Birren, B., Berger, B. & Lander, E. S. Methods in comparative genomics: genome correspondence, gene identification, regulatory motif discovery. *J. Comput. Biol.* (in the press).
44. Turner, R. J., Lovato, M. & Schimmel, P. One of two genes encoding glycyl-tRNA synthetase in *Saccharomyces cerevisiae* provides mitochondrial and cytoplasmic functions. *J. Biol. Chem.* **275**, 27681–27688 (2000).
45. Kurtzman, C. P. & Robnett, C. J. Phylogenetic relationships among yeasts of the ‘Saccharomyces complex’ determined from multigene sequence analyses. *FEMS Yeast Res.* **3**, 417–432 (2003).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank I. Roberts for providing strains; M. Endrizzi, L.-J. Ma, D. Qi and the staff of the MIT/Whitehead Institute Center for Genome Research who generated the shotgun sequence from *Kluyveromyces waltii*; J. Butler and the Arachne assembly team who generated the genome assembly; D. Bartel, S. Calvo, J. Galagan, D. Jaffe and J. Vinson for discussions and comments on the manuscript; and G. Fink and A. Murray for discussions and advice.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.K. (manoli@mit.edu) and E.S.L. (lander@broad.mit.edu). The GenBank accession number for *K. waltii* is AADM01000000.

The star-formation history of the Universe from the stellar populations of nearby galaxies

Alan Heavens¹, Benjamin Panter¹, Raul Jimenez² & James Dunlop¹

¹Institute for Astronomy, University of Edinburgh, Blackford Hill, Edinburgh EH9 3HJ, UK

²Department of Physics and Astronomy, University of Pennsylvania, 209 South 33rd Street, Philadelphia, Pennsylvania 19104-6396, USA

The determination of the star-formation history of the Universe is a key goal of modern cosmology, as it is crucial to our understanding of how galactic structures form and evolve. Observations^{1–12} of young stars in distant galaxies at different times in the past have indicated that the stellar birthrate peaked some eight billion years ago before declining by a factor of around ten to its present value. Here we report an analysis of the ‘fossil record’ of the current stellar populations of 96,545 nearby galaxies, from which we obtained a complete star-formation history. Our results broadly support those derived from high-redshift galaxies. We find, however, that the peak of star formation was more recent—around five billion years ago. We also show that the bigger the stellar mass of the galaxy, the earlier the stars were formed, which indicates that high- and low-mass galaxies have very different histories.

The optical stellar spectrum of a galaxy can be used as a probe of both its past star-formation history, and the metallicity of its gas as a function of time. The spectra of a large sample of nearby galaxies therefore acts as a useful fossil record of the star-formation rate of the Universe, allowing estimates of the star-formation rate over a very wide range of cosmic time, from the same internally consistent data set, and with very small statistical errors. The alternative is to use the finite speed of light to view galaxies at different cosmic epochs, and to look for signs of recent star formation. The present method decouples the star formation from the mass assembly—it takes into account all stars which end up in normal galaxies today, and it is much less sensitive to the large corrections, for extinction and for unobserved small galaxies, which must be applied to high-redshift studies.

The spectral data used in this analysis come from the Sloan Digital Sky Survey data release 1 (SDSS DR1), from an area of 1,360 square degrees. The main sample has red apparent magnitude limits of $15.0 \leq m_R \leq 17.77$, and we also place a limit on surface brightness of $\mu_R < 23.0$ (see ref. 13 for discussion of this). The redshift range is $0.005 < z < 0.34$, with a median of 0.1. This leaves 96,545 galaxies in this study. Full details of the SDSS are available at <http://www.sdss.org/>. The spectra are top-hat smoothed to 20-Å resolution, for comparison with the models of ref. 14, and emission-line regions are removed, as these are generally non-stellar and outside the scope of the theoretical model employed.

The speed of MOPED (see below) means we are not restricted to assuming simple parametrizations of the star-formation history^{15,16}. For each spectrum, we recover the mass of stars created in 11 time periods, which are mostly equally spaced logarithmically in look-back time, separated by factors of 2.07, but with the first boundary at a redshift of two. We assume a Salpeter initial mass function, and a cosmology given by the best-fitting parameters determined by the WMAP satellite¹⁷: $\Omega_m = 0.27$, $\Omega_v = 0.73$, $H_0 = 71 \text{ km s}^{-1} \text{ Mpc}^{-1}$. For each time period, we also recover the average metallicity of the gas, and an overall dust parameter for the galaxy, assuming a Large Magellanic Cloud extinction curve¹⁸. Thus we have a 23-dimensional parameter space to search. A straightforward maximum-likelihood solution using the full spectrum of all the SDSS galaxies is

impractical, so we use the patented radical lossless data compression algorithm MOPED¹⁹ to compress each spectrum to 23 numbers, and we find the set of 23 parameters which fit these MOPED coefficients most accurately. The massive data compression allows much faster determination of the parameters and the errors on them, and, importantly, can be shown to be lossless in ideal cases, in the sense that the error bars are not increased by using the MOPED coefficients rather than the entire spectrum. More details are given in papers developing and testing MOPED^{19–21}.

With this method, we obtain the mass of stars created in each galaxy g in each period of time (relative to the time of emission of the galaxy light), $\delta M_{\star, g}(t)$. After redistribution of the stars to a set of time bins fixed in cosmic time, we estimate the star-formation rate per unit co-moving volume by $\dot{\rho}_{\star}(t) = \sum_g \delta M_{\star, g} / (V_{\text{max}} \delta t)$, where δt is the width of the time bin. V_{max} is the maximum volume in which the galaxy could be placed and still contribute to the star-formation rate estimate at time t . We have no knowledge of the star-formation rate at redshifts less than the observed redshift of the galaxy, so this introduces a redshift cutoff. Other limits come from the magnitude and surface brightness limits of the sample. These are calculated by computing the galaxy spectrum as a function of time (from the recovered star-formation-rate parameters) and computing the expected observed R -band flux and surface brightness. We take account of the three-arcsec fibre aperture by correcting the spectrum upwards by the ratio of the flux in R in a petrosian radius (determined by the photometry) to the fibre flux. For individual

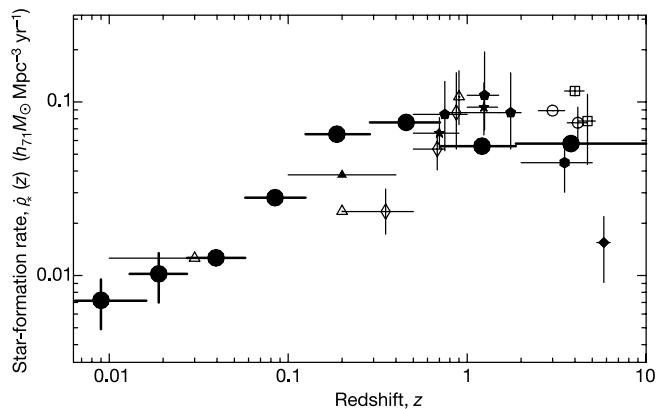


Figure 1 The star-formation history of the Universe. The star-formation rate recovered from the ‘fossil record’ in the SDSS is shown by the eight large filled circles. The horizontal error bars represent the size of the bin in redshift. Vertical error bars represent bootstrap errors, and are invisibly small for most bins. Also shown are independent determinations using instantaneous measurements of the star-formation rate, as follows: H α measurements (open triangles) at $z \approx 0.03$ (ref. 1), $z \approx 0.2$ (ref. 5) and $z \approx 0.9$ (ref. 6); ultraviolet from Subaru¹¹ (open squares), GOODS¹² (filled diamond), HST and so on³ (open circles), CFRS² (open diamonds), HDF⁴ (filled pentagons), galaxies⁷ (star shapes) and galaxies⁸ (filled triangle). The filled hexagon at $z = 3.5$ represents a new estimate of the star-formation density provided by the sub-mm galaxies in the redshift range $2 < z < 5$. This was derived by integrating the sub-mm source number counts⁹ down to $S_{850 \mu\text{m}} = 1 \text{ mJy}$, and assuming that 75% of such sources lie at $z > 2$ (in line with recent redshift measurements¹⁹). In general, the agreement is very good. However, there are two important results in our study which result from the extremely small vertical error bars. First, we find that 26% of the mass of stars in the present-day Universe were formed at $z > 2$. Second, while we confirm the previous measurements of a generally high level of star-formation activity at $z \approx 1$, we find that global star-formation density peaked at significantly lower redshifts than previously claimed, in the bin spanning the redshift range $0.3 < z < 0.8$. The reason for this difference is clarified in Fig. 2. h_{71} is the Hubble constant in units of $71 \text{ km s}^{-1} \text{ Mpc}^{-1}$.

galaxies this is likely to fail, but for the population as a whole there is evidence from the colours¹⁶ that fibre-placing is such that there is no significant systematic difference in the sampling of stellar populations by the spectroscopy and photometry.

In Fig. 1 we show the co-moving star-formation rate determined by MOPED and SDSS DR1, as a function of redshift, along with results from other determinations, largely based on instantaneous star-formation rate indicators (ultraviolet flux, H α emission, sub-millimetre (mm) emission and so on). At a basic level, the new data show good agreement over the redshift range of $0.01 < z < 6$. The low-redshift decline of the star-formation rate is clearly seen, and the rate also appears to have flattened. Moreover, the level of star-formation activity inferred at high redshift ($z > 2$) is in rather good agreement with the completely independent values inferred from observations of high-redshift galaxies, suggesting that the dust corrections made in such studies are reasonable, if a little over-estimated. The broad agreement of the star-formation rate determined locally from the fossil record of the SDSS with the determinations from high-redshift galaxies gives support for the copernican principle: that is, that the Earth has no special location in the Universe. It also supports the current standard cosmological model, as the volume elements assumed in the high-redshift studies are sensitive to the cosmological parameters. Note, however, that this does assume that our assumption about the initial mass function is reasonable, as we are inferring the number of early-forming high-mass stars (now dead) from the numbers of early-forming low-mass stars, which are still contributing to the galaxy spectrum.

One of our main results is that the period of star formation is broader than previously thought, and occurs at a lower redshift $z \approx 0.6$, rather than one or more. Specifically, we find that 26% of

the mass of stars in the present-day Universe was formed at $z > 2$ (compare ref. 22). The average metallicity rises from 0.44 (relative to the solar value) at high z to a peak of 0.8 at $z \approx 1$ before declining to a level around 0.25 at the present time.

As we explain below, we believe our result differs because it includes the contributions made by all galaxies over a very wide mass range, extending down to galaxies with $L \approx 2 \times 10^{-3} L_*$, where L_* is the characteristic luminosity of a large galaxy. Note that virtually all 96,545 galaxies contribute to the $z > 0.3$ bins, so the statistical errors (bootstrap estimates) are negligible in comparison with modelling uncertainties and residual uncertainties in flux calibration. We also note that because we are not dominated by statistical errors, the errors are smaller in this approach than by analysing, for example, some appropriately-weighted average of the spectra themselves. Supplementary Fig. 1 shows that parameter recovery is robust.

Our second major result is apparent when we present the star-formation history of the sample, divided into different ranges of observed stellar mass. We see clearly in Fig. 2 that the redshift at which star-formation activity peaks is an essentially monotonically increasing function of final stellar mass. Star-formation activity in the galaxies in the lowest-mass bins ($M_* \approx 10^{10} M_\odot$, where M_* is the stellar mass and $M_\odot$ the solar mass) peaked at $z \approx 0.2$, whereas for galaxies that are an order of magnitude more massive the peak lies at $z \approx 0.5$. At still higher masses, galaxies with masses comparable to a present-day L_* galaxy appear to have experienced a peak in activity at $z \approx 0.8$, whereas the highest-mass systems ($M_* > 10^{12} M_\odot$) show a monotonic decline in star-formation rate in our data, with any peak constrained to lie at $z > 2$. This provides a natural explanation for why the most massive star-forming systems, such as the selected luminous, sub-mm galaxies, should largely be found to lie at high redshifts ($z > 2$; ref. 10), while at the same time providing further evidence that the bright sub-mm galaxies are indeed the progenitors of today's massive ellipticals²³. The importance of low-mass systems in low-redshift star formation has been noted (see, for example, refs 24 and 25).

Indeed, the strong mass-dependence of the star-formation history also provides a natural explanation of the high redshift of peak star-formation activity seen in other surveys, because they are sensitive only to the most massive objects. That we have now discovered that global star-formation activity in fact peaks at rather a modest redshift is because the peak epoch of star-formation activity in objects of lower (present-day stellar) masses ($< 3 \times 10^{11} M_\odot$) was at $z \leq 0.5$, and because such lower-mass galaxies make a significant contribution to the overall star-formation density. Given present-day observational capabilities, this result could only be revealed by a method such as used here, because the fossil record approach allows us to explore the star-formation history of galaxies spanning over two decades in mass. Supplementary Fig. 2 demonstrates consistent results from volume-limited subsamples.

Finally, we note that the mass dependence of peak star-formation epoch revealed in Fig. 2 appears to mirror the mass dependence of black-hole activity as recently seen in redshift surveys of both radio-selected²⁶ and X-ray-selected²⁷ active galactic nuclei. Such apparently anti-hierarchical behaviour ('downsizing') is in fact quite consistent with the standard cosmological model, in which galaxies form in small units and merge—our method makes no statement about whether the stellar mass at high redshift was in smaller units or not. The behaviour we see is based on the present-day stellar mass of the galaxies, and in general we would expect more massive systems to be part of large-scale overdensities, in which first star formation would occur earlier. Furthermore, these results suggest a very different formation history for low- and high-mass systems, because the mass assembly proceeds in the opposite direction to the star formation. (This will be explored in a separate paper.)

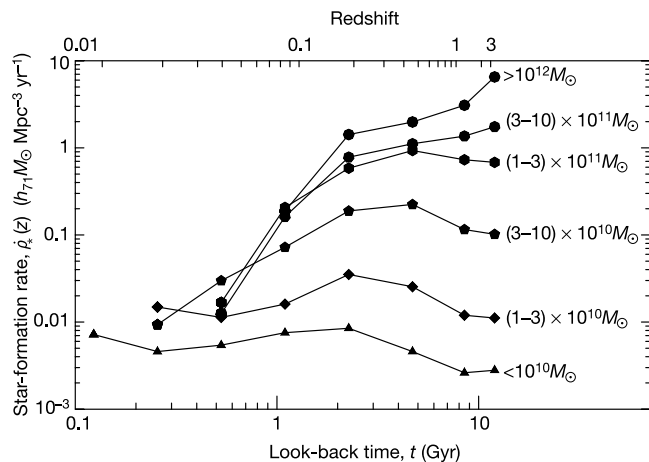


Figure 2 The star-formation rate as a function of the observed stellar mass of the galaxy. For clarity, the curves are offset vertically, successively by 0.5 log units, except for the most massive galaxies, which are offset by an additional 1.0. Note the clear trend for galaxies with larger present-day stellar mass to have formed their stars earlier. The bulk of the star-formation rate at $z \approx 0.5$ comes from galaxies with present-day stellar masses in the range $3 - 30 \times 10^{10} M_\odot$. Note that the graph makes no statement about when the masses were aggregated. To test further the robustness of our findings we have reconstructed the star-formation history, changing the dust model and the theoretical stellar population models. The shape of the star-formation history is hardly changed by changing the dust model to the Calzetti extinction law²⁸, or by using Bruzual–Charlot²⁹ stellar population models; the latter allows us to assess the effect of systematic errors in our determination of the star-formation rate, because the Jimenez and Bruzual–Charlot models are based on different stellar interior and atmosphere models.

Received 19 December 2003; accepted 9 March 2004; doi:10.1038/nature02474.

- Gallego, J., Zamorano, J., Aragon-Salamanca, A. & Rego, M. The current star formation rate of the local Universe. *Astrophys. J. Lett.* **455**, 1–4 (1995).
- Lilly, S. J., Le Fevre, O., Hammer, F. & Crampton, D. The Canada-France redshift survey: The luminosity density and star formation history of the Universe to $z \sim 1$. *Astrophys. J. Lett.* **460**, 1–4 (1996).
- Steidel, C. C., Giallisco, M., Pettini, M., Dickinson, M. & Adelberger, K. L. Spectroscopic confirmation of a population of normal star-forming galaxies at redshifts $z > 3$. *Astrophys. J. Lett.* **462**, 17–21 (1996).
- Connolly, A. J., Szalay, A. S., Dickinson, M., Subbarao, M. U. & Brunner, R. J. The evolution of the global star formation history as measured from the Hubble Deep Field. *Astrophys. J. Lett.* **486**, 11–14 (1997).
- Tresse, L. & Maddox, S. J. The $H\alpha$ luminosity function and star formation rate at $z \sim 0.2$. *Astrophys. J.* **495**, 691–697 (1998).
- Glazebrook, K., Blake, C., Economou, F., Lilly, S. & Colless, M. Measurement of the star formation rate from $H\alpha$ in field galaxies at $z = 1$. *Mon. Not. R. Astron. Soc.* **306**, 843–856 (1999).
- Cowie, L., Songaila, A. & Barger, A. J. Evidence for a gradual decline in the universal rest-frame ultraviolet luminosity density for $z < 1$. *Astron. J.* **117**, 2656–2665 (1999).
- Sullivan, M. et al. An ultraviolet-selected galaxy redshift survey—II. The physical nature of star formation in an enlarged sample. *Mon. Not. R. Astron. Soc.* **312**, 442–464 (2000).
- Scott, S. et al. The SCUBA 8-mJy survey: I—Sub-mm maps, sources and source counts. *Mon. Not. R. Astron. Soc.* **331**, 817–838 (2002).
- Chapman, S. C., Blain, A. W., Ivison, R. J. & Smail, I. R. A median redshift of 2.4 for galaxies bright at submillimetre wavelengths. *Nature* **422**, 695–698 (2003).
- Ouchi, M. et al. A census of Lyman break galaxies at $z = 4$ and 5 in the Subaru Deep Fields: Photometric properties. Preprint at (<http://www.arXiv.org/astro-ph/0309657>) (2003).
- Stanway, E. R., Bunker, A. J. & McMahon, R. G. Lyman break galaxies and the star formation rate of the Universe at $z \sim 6$. *Mon. Not. R. Astron. Soc.* **342**, 439–445 (2003).
- Shen, S. et al. The size distribution of galaxies in the Sloan Digital Sky Survey. *Mon. Not. R. Astron. Soc.* **343**, 978–994 (2003).
- Jimenez, R., MacDonald, J., Dunlop, J. S., Padoan, P. & Peacock, J. A. Synthetic stellar populations: single stellar populations, stellar interior models and primordial protogalaxies. *Mon. Not. R. Astron. Soc.* **349**, 240–254 (2004).
- Baldry, I. K. et al. The 2dF galaxy redshift survey: Constraints on cosmic star formation history from the cosmic spectrum. *Astrophys. J.* **569**, 582–594 (2002).
- Glazebrook, K. et al. The Sloan Digital Sky Survey: The cosmic spectrum and star formation history. *Astrophys. J.* **587**, 55–70 (2003).
- Spergel, D. N. et al. First year Wilkinson Microwave Anisotropy Probe (WMAP) observations: Determination of cosmological parameters. *Astrophys. J. Suppl.* **148**, 175–194 (2003).
- Gordon, K. D., Clayton, G. C., Misselt, K. A., Landolt, A. U. & Wolff, M. J. A quantitative comparison of the Small Magellanic Cloud, Large Magellanic Cloud, and Milky Way ultraviolet to near-infrared extinction curves. *Astrophys. J.* **594**, 279–293 (2003).
- Heavens, A. F., Jimenez, R. & Lahav, O. Massive lossless data compression and multiple parameter estimation from galaxy spectra. *Mon. Not. R. Astron. Soc.* **317**, 965–972 (2000).
- Reichardt, C., Jimenez, R. & Heavens, A. F. Recovering physical parameters from galaxy spectra using MOPED. *Mon. Not. R. Astron. Soc.* **327**, 849–867 (2001).
- Panther, B., Heavens, A. F. & Jimenez, R. Star formation and metallicity history of the SDSS galaxy survey: unlocking the fossil record. *Mon. Not. R. Astron. Soc.* **343**, 1145–1154 (2003).
- Dickinson, M., Papovich, C., Ferguson, H. C. & Budavari, T. The evolution of the global stellar mass density at $0 < z < 3$. *Astrophys. J.* **587**, 25–40 (2003).
- Dunlop, J. S. in *Deep Millimetre Surveys* (eds Lowenthal, J. D. & Hughes, D. H.) 11–18 (World Scientific, Singapore, 2001).
- Perez-Gonzalez, P. G. et al. Stellar populations in local star-forming galaxies—II. Recent star formation properties and stellar masses. *Mon. Not. R. Astron. Soc.* **338**, 525–543 (2003).
- Fujita, S. S. et al. The $H\alpha$ luminosity function and star formation rate at $z \sim 0.24$ based on Subaru deep imaging data. *Astrophys. J. Lett.* **586**, 115–118 (2003).
- Waddington, I., Dunlop, J. S., Windhorst, R. A. & Peacock, J. A. The LBDS Hercules sample of millijansky radio sources at 1.4 GHz: II. Redshift distribution and radio luminosity function. *Mon. Not. R. Astron. Soc.* **328**, 882–896 (2001).
- Hasinger, G. The X-ray background and AGNs. in *The Restless High Energy Universe* (eds van den Hevel, E. P. J., in't Zand, J. J. M. & Wijers, R. A. M. J.) (Nucl. Phys. B. Suppl. Ser. in the press); preprint at (<http://www.arXiv.org/astro-ph/0310804>) (2003).
- Calzetti, D. Reddening and star formation in starburst galaxies. *Astron. J.* **113**, 162–184 (1997).
- Bruzal, G. & Charlot, S. Spectral evolution of stellar populations using isochrone synthesis. *Astrophys. J.* **405**, 538–553 (1993).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We are grateful to M. Pettini and M. Ouchi for helpful remarks. The SDSS is managed by the Astrophysical Research Consortium (ARC) for the Participating Institutions. The participating institutions are The University of Chicago, Fermilab, the Institute for Advanced Study, the Japan Participation Group, The Johns Hopkins University, Los Alamos National Laboratory, the Max Planck Institute for Astronomy (MPIA), the Max Planck Institute for Astrophysics (MPA), New Mexico State University, the University of Pittsburgh, Princeton University, the United States Naval Observatory and the University of Washington.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be sent to A.H. (afh@roe.ac.uk).

Perennial water ice identified in the south polar cap of Mars

Jean-Pierre Bibring¹, Yves Langevin¹, François Poulet¹, Aline Gendrin¹, Brigitte Gondet¹, Michel Berthé¹, Alain Soufflot¹, Pierre Drossart², Michel Combes², Giancarlo Bellucci³, Vassili Moroz⁴, Nicolas Mangold⁵, Bernard Schmitt⁶ & the OMEGA team*

¹Institut d'Astrophysique Spatiale, Orsay Campus, 91405, France

²LESIA, Observatoire de Paris/Meudon, 92195 Meudon, France

³IFSI-INAf, Rome, Italy

⁴IKI, Moscow, Russia

⁵OrsayTerre, Orsay Campus, 91405, France

⁶Laboratoire de Planétologie de Grenoble, 38400, France

* A list of all the members of the OMEGA team and their affiliations appears at the end of the paper

The inventory of water and carbon dioxide reservoirs on Mars are important clues for understanding the geological, climatic and potentially exobiological evolution of the planet¹. From the early mapping observation of the permanent ice caps on the martian poles^{2,3}, the northern cap was believed to be mainly composed of water ice, whereas the southern cap was thought to be constituted of carbon dioxide ice. However, recent missions (NASA missions Mars Global Surveyor and Odyssey) have revealed surface structures⁴, altimetry profiles⁵, underlying buried hydrogen⁶, and temperatures of the south polar regions that are thermodynamically consistent with a mixture of surface water ice and carbon dioxide⁷. Here we present the first direct identification and mapping of both carbon dioxide and water ice in the martian high southern latitudes, at a resolution of 2 km, during the local summer, when the extent of the polar ice is at its minimum. We observe that this south polar cap contains perennial water ice in extended areas: as a small admixture to carbon dioxide in the bright regions; associated with dust, without carbon dioxide, at the edges of this bright cap; and, unexpectedly, in large areas tens of kilometres away from the bright cap.

The ESA/Mars Express Orbiter⁸ was inserted into Mars orbit on 25 December 2003. The Observatoire pour la Minéralogie, l'Eau, les Glaces et l'Activité (OMEGA) instrument⁹, one of the seven investigating instruments on board, is an imaging spectrometer analysing the diffused solar light and the planetary thermal emission. On each resolved pixel, 1.2 mrad in the instantaneous field of view, OMEGA acquires a spectrum in 352 contiguous spectral channels from 0.35 to 5.1 μm , with a spectral sampling ranging from 7 nm (in the visible) to 13 nm (from 1.0 to 2.7 μm) and 20 nm (from 2.7 to 5.1 μm). A few initial observations were performed soon after Mars orbit insertion, in particular while the spacecraft was overflying the south polar areas. The altitude of observation ranged from $\sim 1,500$ km to $\sim 2,000$ km, providing an OMEGA surface sampling of ~ 2 km. OMEGA thus mapped a large fraction of the south polar regions, along four distinct orbits, from 18 January to 11 February 2004 (solar longitude, $L_{\odot} = 335^{\circ}$ to 348°)—that is, about one month before the martian southern autumn equinox. At the time of the observations, the Sun's elevation was very low ($< 10^{\circ}$); however, given the very high performances of OMEGA, several tens of thousands of spectra were acquired with signal-to-noise ratios of > 100 over almost the entire spectral domain. From the acquired spectra, it is possible to derive coupled maps of a variety of parameters and properties: in particular, the major icy constituents, CO_2 and H_2O , have several unambiguous diagnostic spectral signatures enabling the mapping of their respective distributions over the imaged areas (see below).

The OMEGA spectral images exhibit well-characterized high albedo perennial polar ice patterns. From their near-infrared spectrum, we can assert that these bright areas are those where

CO₂ ice is highly concentrated (Fig. 1a).

A crucial finding is the identification of H₂O ice, with a varying concentration over a much larger area (Fig. 1b). H₂O ice is found in three distinct units, as follows.

Unit 1. On the bright cap. In this unit, water ice is mixed with large concentrations of CO₂ ice. This makes its spectral identification more difficult than in the two other units, as CO₂ ice exhibits spectral features in the same spectral domain. It requires a careful

analysis, illustrated in Fig. 2, where a typical OMEGA spectrum of a pixel located in this CO₂ ice cap is compared to modelled spectra obtained with either pure CO₂ ice or a mixture of H₂O ice and CO₂ ice. We show that pure CO₂ ice cannot account for the observed spectra: the presence of H₂O ice is required. The best fit is obtained with a mixture of 15 wt% H₂O ice, in the form of an intimate molecular mixture, rather than a simple geometrical mixture of two pure components.

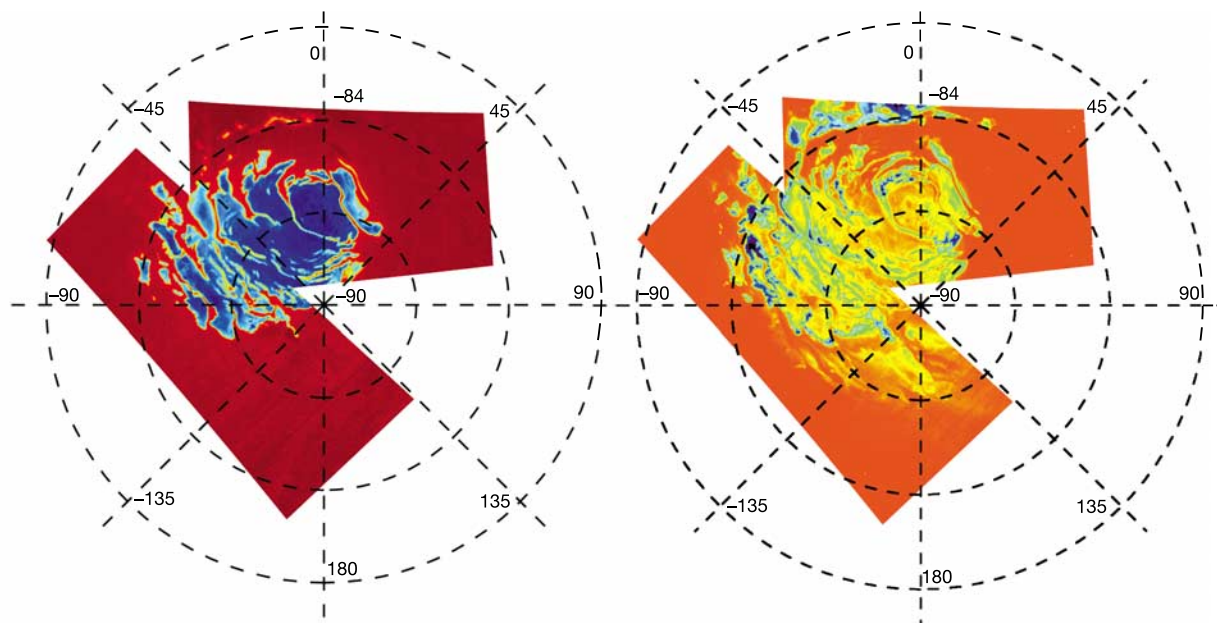


Figure 1 Global maps of CO₂ and H₂O ices at the south pole of Mars. Left, the CO₂-ice absorption feature is scaled from blue (deep) to brown (CO₂-ice-free areas); right, mapping of the H₂O ice, from blue (deep absorption) to red (ice-free). Comparison shows

that the H₂O-ice areas extend far beyond the CO₂-rich bright cap, along its scarps up to isolated units tens of kilometres wide.

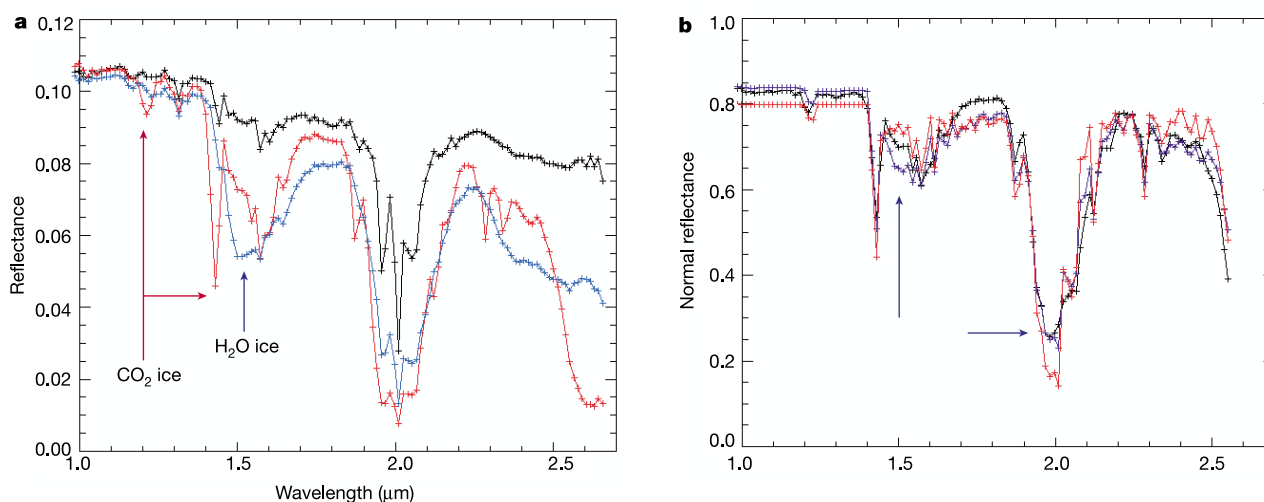


Figure 2 Near-infrared spectra of CO₂ ice and H₂O ice in Mars polar units. **a**, 1.0 to 2.6 μm part of spectra acquired by OMEGA on an ice-free zone (black), on the CO₂-ice-rich bright cap (red) and on an H₂O-ice-rich scarp (blue). In addition to the major atmospheric spectral features, they exhibit the specific CO₂ and H₂O ice absorptions, some of them being indicated (arrows). The scarps are seen to have none of the numerous CO₂-ice diagnostic features. **b**, OMEGA spectrum of a pixel located on the CO₂-rich cap (black), after removal of the atmospheric contribution, is compared to the best-fit modelled spectrum of a sample containing CO₂ ice only (red, r.m.s. = 0.057) and to the

spectrum of a mixture of CO₂ and H₂O (blue, r.m.s. = 0.035), indicating that some H₂O is required to account for the pattern observed at 1.5 and 2.0 μm (arrows). The H₂O/CO₂ ratio providing the best fit is 15 wt% H₂O, in the form of an intimate mixing. However, some departure from the actual spectrum is still present, mainly at 1.5 μm ; this could be accounted for by a complex mixture including H₂O/CO₂ clathrate, which could not yet be simulated in the absence of relevant optical constants. The synthetic spectra have been calculated using the Shkuratov–Poulet¹⁵ model.

Unit 2. On the scarps around the residual cap. This H₂O ice is almost completely CO₂-ice-free, and appears darker than the surrounding bright CO₂-ice-rich cap; Fig. 3 illustrates the correlation between the albedo, the H₂O content and the CO₂ content of the ice: the darker the ice, the richer it is in H₂O, and the more depleted it is in CO₂.

Unit 3. Along vast zones expanding down slope in stratified terrains, tens of kilometres wide, and tens of kilometres away from

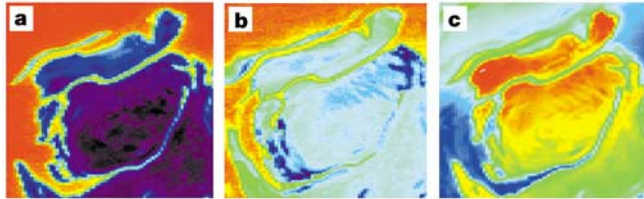


Figure 3 Correlated variations of the CO₂ ice, H₂O ice and albedo. **a**, Map of the CO₂ ice, from deep blue (higher concentration) to red (ice-free), matching the bright cap, centred around 87° (latitude) and 0° (longitude). **b**, Same area mapped for its H₂O-ice content. **c**, Same area, seen in its near infrared I/F (intensity divided by solar flux), from dark (blue to green) to bright (red). Local I/F variations within the cap are clearly correlated to the H₂O/CO₂ ratio: the bluest water-ice areas are highly depleted in CO₂, and darker, while the brighter areas are those where H₂O is depleted in favour of CO₂.

the CO₂-ice-rich perennial caps. These water-ice-rich areas are CO₂-free: they appear much darker than the bright CO₂-ice-rich cap, both in the visible and the near-infrared, which reflects their high dust concentration, and would confirm the scenarios where H₂O precipitates together with dust. The OMEGA processed data show that the H₂O-to-dust ratio varies, with no clear correlation with the local topography or relief (Sun illumination angle).

The OMEGA units 2 and 3 partly match the Af unit in the USGS I-2686 geological map¹⁰. Af is defined on the basis of its albedo and colour, found to be “intermediate between those of residual polar ice cap and dust unit”¹⁰. The Af unit is interpreted as a “mixture of CO₂ frost and defrosted ground”¹⁰, while the OMEGA observation shows CO₂-ice-free exposed mixtures of water ice and dust.

Figure 4 illustrates the relation found between the composition as inferred from the OMEGA spectral images and the surface morphology revealed by the Mars Orbiting Camera (MOC) high-resolution images on the Mars Global Surveyor (MGS) mission. From right to left of the MOC strip (Fig. 4d), we observe the bright polar ice, which exhibits irregular patterns similar to ‘Swiss cheese’ (ref. 11); OMEGA results demonstrate that this area is predominantly constituted of CO₂ ice. It is followed by the H₂O-ice-rich areas, free of CO₂, which by contrast are characterized by very smooth surfaces. Within these H₂O-ice-rich areas, close to the CO₂ ice, we observe polygonal features; this structure is consistent with

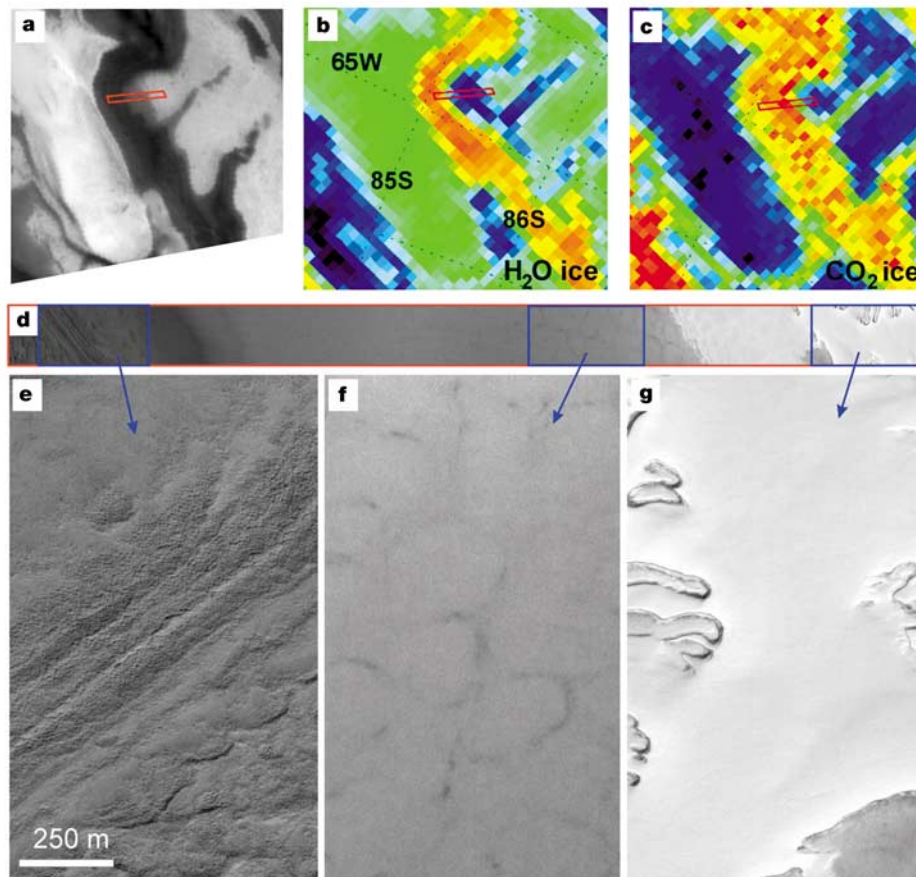


Figure 4 Transition from CO₂ ice to H₂O ice. **a**, MOC (NASA/JPL/MSSS) context image, exhibiting the bright polar areas. **b**, Map, within the same area, of the H₂O ice as identified by OMEGA (processed 1.5 μ m feature, see Fig. 2), with a blue to red colour scale; the bluest is the richest in H₂O. **c**, Same mapping of the CO₂-ice features. **d**, Enlargement of the strip (image M13-01891, at $L_{\odot} = 325.3^{\circ}$) marked by the red rectangles in **a**, **b** and **c**. It exhibits the variety of terrains identified by the MOC, examples being enlarged

following the arrows (**e**, **f**, **g**), from the bright ‘Swiss cheese’-like structured ice, towards the dark layer terrain on the left. The OMEGA-derived maps indicate the composition of each unit: CO₂-rich ice for the bright unit; ice-free for the layer terrains; and H₂O ice for the smooth dark terrains in between, with polygonal structures closer to the bright cap. These structures can also be observed in the transition zone (between **f** and **g**), through a thin covering layer of CO₂ ice.

them resulting from thermal retraction. At the transition between the bright cap and the smooth dusty (darker) H₂O-ice-exposed units, these polygonal features are seen through a thin and bright CO₂ ice layer. The ice-free stratified terrains expand then to the left of the image.

Wherever CO₂ ice is stable, it traps solid H₂O. OMEGA data show large areas where CO₂ ice has sublimated away, while water ice remains, mostly mixed with dust. The perennial southern cap thus by far exceeds the bright CO₂-rich icy area. Moreover, the lack of slope change at the transition between the bright cap and the surrounding water-ice polygonal areas is an indication that the CO₂-ice layer might well be restricted to a fairly thin layer, no more than some metres in depth, in agreement with previous estimates derived from shadow measurements in quasi-circular depressions¹¹. OMEGA spectral images of the bright cap and of its scarps are consistent with a model in which H₂O ice is present in the underlying material, a scenario which has been proposed to explain the morphology of the 'Swiss cheese' terrains¹².

The underlying material, over 1–3 km in thickness and some 400 km in size, according to the Mars Orbiter Laser Altimeter (MOLA) polar cap profile, is likely to constitute a distinct geological unit as inferred from its morphology¹¹. The direct OMEGA surface compositional measurements support a model in which the bulk of this unit is constituted of dust mixed with water ice, as predicted by mechanical modelling¹³. The perennial south polar cap would then constitute a significant fraction of the overall H₂O reservoir. Conversely, if CO₂ ice is indeed restricted to the bright areas, possibly concentrated in a thin surface veneer, it would not constitute a major sink for the initial atmospheric CO₂.

The inventory of condensed volatiles on Mars, in all potential phases (gas, clouds, hydrated minerals, frosts, ices, permafrost, liquids), at and below its surface, is central to understanding the evolution of the planet, from geological timescales to seasonal variations. Part of the answer to the question of Mars having harboured life in the past, and to the issue of Mars hosting future human exploration, is to be deciphered in this inventory. This is why contributing to this inventory, in particular on the poles, remains a major goal for Mars space exploration¹⁴. With the ongoing Mars Express mission just starting its global coverage of Mars, OMEGA will continue mapping the CO₂ and H₂O condensation/sublimation cycles at all south and north latitudes; the goal is to provide a quantitative evaluation of the seasonal and perennial condensed reservoirs of these crucial volatile species. In parallel, the Mars Advanced Radar for Subsurface and Ionosphere Sounding (MARSIS) experiment will start its mapping phase in May 2004. It should enable an accurate determination of the volume of the H₂O-ice-rich surface permafrost identified by OMEGA at the south pole, and compare it with a potential subsurface permafrost considered up until now to be the major Mars water reservoir. □

Received 19 February; accepted 5 March 2004; doi:10.1038/nature02461.
Published online 17 March 2004.

- Kieffer, H. H. & Zent, A. A. *Mars* (eds Kieffer, H. H., Jakovsky, B. M., Snyder, C. W. & Matthews, M. S.) 1180–1220 (Univ. Arizona Press, Tucson/London, 1992).
- Kieffer, H. H. Mars south polar spring and summer temperatures: a residual CO₂ frost. *J. Geophys. Res.* **84**, 8263–8288 (1979).
- Paige, D. A., Herkenhoff, K. E. & Murray, B. C. Mariner 9 observations of the south polar cap of Mars: evidence of residual CO₂ frost. *J. Geophys. Res.* **95**, 1319–1335 (1990).
- Smith, P. H. *et al.* The global topography of Mars and implications for surface evolution. *Science* **284**, 1495–1503 (1999).
- Schenk, P. M. & Moore, J. M. Stereo topography of the south polar region of Mars: Volatile inventory and Mars Polar Lander landing site. *J. Geophys. Res.* **E 105**, 24529–24546 (2000).
- Boynton, W. V. *et al.* Distribution of hydrogen in the near surface of Mars: Evidence for subsurface ice deposits. *Science* **297**, 81–85 (2002).
- Titus, N., Kieffer, H. H. & Christensen, P. R. Exposed water ice discovered near the South Pole of Mars. *Science* **299**, 1048–1051 (2003).
- Mars Express Project Team. Mars Express: closing in on the Red Planet. *ESA Bull.* **115**, 10–17 (2003).
- Bibring, J.-P. *et al.* OMEGA: Observatoire pour la Minéralogie, l'Eau, les Glaces et l'Activité, (SP-1240, ESA, in the press).
- Herkenhoff, K. E. *Geologic Map of the MTM-85000 Quadrangle, Planum Australe Region of Mars* (USGS Geologic Investigations Map Series I-2686, Washington, DC, 2001).

- Thomas, P. C. *et al.* North–south geological differences between the residual polar caps on Mars. *Nature* **404**, 161–164 (2000).
- Byrne, S. & Ingersoll, A. P. A sublimation model for martian south polar ice features. *Science* **299**, 1051–1053 (2003).
- Nye, J. E., Durham, W. B., Schenk, P. M. & Moore, J. M. The instability of a south polar cap on Mars composed of carbon dioxide. *Icarus* **144**, 449–455 (2000).
- Clifford, S. M. *et al.* The state and future of Mars polar science and exploration. *Icarus* **144**, 210–242 (2000).
- Poulet, F., Cuzzi, J. N., Cruikshank, D. P., Roush, T. & Dalle Ore, C. M. Comparison between the Shkuratov and Hapke scattering theories for solid planetary surfaces: Application to the surface composition of two Centaurs. *Icarus* **160**, 313–324 (2002).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.-P.B. (bibring@ias.u-psud.fr).

The OMEGA team M.Berthé¹, J.-P.Bibring¹, Y.Langevin¹, S.Erard¹, O.Forni¹, A.Gendrin¹, B.Gondet¹, N.Manaud¹, F.Poulet¹, G.Poulléau¹, A.Soufflot¹, M.Combes², P.Drossart², T.Encrenaz², T.Fouchet², R.Melchiorri², G.Bellucci³, F.Altieri³, V.Formisano³, G.Bonello³, S.Fonti³, F.Capaccioni³, P.Cerroni³, A.Coradini³, V.Kottsov⁴, N.Ignatiev⁴, V.Moroz⁴, D.Titov⁴, L.Zasova⁴, N.Mangold⁵, P.Pinot⁶, B.Schmitt⁶, C.Sotin⁶, E.Hauber⁹, H.Hoffmann⁹, R.Jaumann⁹, U.Keller¹⁰, R.Arvidson¹¹, J.Mustard¹², T.Duxbury¹³ & F.Forget¹⁴

Affiliations for authors: 1, IAS, Orsay Campus, 91405, France; 2, LESIA, Observatoire de Paris/Meudon, 92195 Meudon, France; 3, IFSI-INAF, Rome, Italy; 4, IKI, Moscow, Russia; 5, OrsayTerre, Orsay Campus, 91405, France; 6, Observatoire Midi-Pyrénées, 31000 Toulouse, France; 7, Laboratoire de Planétologie, 38400, France; 8, Planétologie, Université de Nantes, France; 9, DLR, Berlin, Germany; 10, MPAE, Lindau, Germany; 11, Earth and Planetary Sciences, Washington University, St Louis, Missouri 63130, USA; 12, Geological Sciences, Brown University, Providence, Rhode Island 02912, USA; 13, JPL, Pasadena, California, USA; 14, LMD, Université de Paris 6, 75252, Paris, France

Electronic reconstruction at an interface between a Mott insulator and a band insulator

Satoshi Okamoto & Andrew J. Millis

Department of Physics, Columbia University 538 West 120th Street, New York, New York 10027, USA

Surface science is an important and well-established branch of materials science involving the study of changes in material properties near a surface or interface. A fundamental issue has been atomic reconstruction: how the surface lattice symmetry differs from the bulk. 'Correlated-electron compounds' are materials in which strong electron–electron and electron–lattice interactions produce new electronic phases, including interaction-induced (Mott) insulators, many forms of spin, charge and orbital ordering, and (presumably) high-transition-temperature superconductivity^{1,2}. Here we propose that the fundamental issue for the new field of correlated-electron surface/interface science is 'electronic reconstruction': how does the surface/interface electronic phase differ from that in the bulk? As a step towards a general understanding of such phenomena, we present a theoretical study of an interface between a strongly correlated Mott insulator and a band insulator. We find dramatic interface-induced electronic reconstructions: in wide parameter

ranges, the near-interface region is metallic and ferromagnetic, whereas the bulk phase on either side is insulating and anti-ferromagnetic. Extending the analysis to a wider range of interfaces and surfaces is a fundamental scientific challenge and may lead to new applications for correlated electron materials.

To assess the effects of a surface or interface on correlated-electron behaviour we require to understand the changes in the parameters governing bulk correlated-electron behaviour. The three key factors are interaction strengths, bandwidths and electron densities^{1,3}, all of which may change near a surface or interface. In most cases, surface- or interface-induced changes in all three factors will contribute, but in developing a general understanding it is desirable first to study the different effects in isolation. Here we focus on the effect of electron-density variation caused by the spreading of charge across an interface. A different charge distribution effect—the compensation of a polar surface by electronic charge rearrangement—was argued to change the behaviour of C_{60} films⁴. (Indeed, Hesper *et al.*⁴ coined the term “electronic reconstruction” in reference to this specific effect; we suggest that this useful phrase be applied more generally to denote electronic phase behaviour that is fundamentally different at a surface from in bulk.) Proximity to a surface or interface can also change the electron interaction parameters^{5,6}, the electron bandwidth⁷, and level degeneracy⁸. Experimental studies of surfaces^{9–12} and heterostructures¹³ have been interpreted along these lines. Our work is inspired by the recent striking experimental results of ref. 14; these authors fabricated atomically precise digital heterostructures by inserting a controllable number of planes of LaTiO_3 (a Mott insulator, that is, a material which is insulating although standard band theory

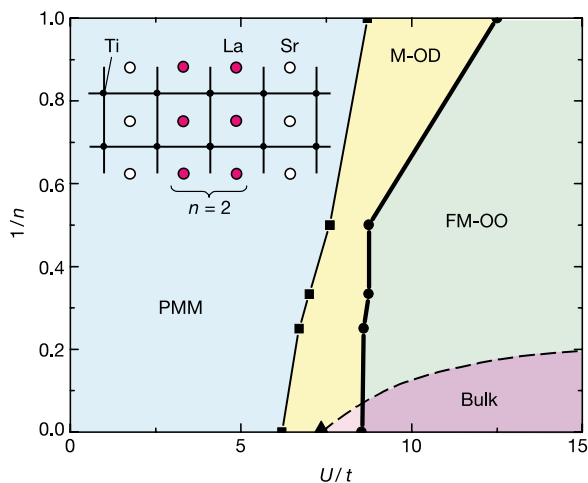


Figure 1 Ground-state phase diagram computed in Hartree–Fock approximation as a function of the on-site Coulomb interaction U and the inverse of the La layer number n . For small U values the ground state is a paramagnetic metal with no orbital ordering (PMM region; shaded blue). The solid line marked with squares denotes a second-order transition to an orbitally disordered magnetic state (M-OD; shaded yellow) that is ferromagnetic for $n = 1$; for $n > 1$ each (001) Ti layer is uniformly polarized, but the magnetization direction alternates from layer to layer, leading to a ferrimagnetic state ($n = 2, 4, 6, \dots$; odd number of occupied Ti layers) or antiferromagnetic state ($n = 3, 5, 7, \dots$; even number of occupied Ti layers). The solid line marked with circles denotes a first-order transition to a fully polarized ferromagnetic state with (OO π) orbital order (FM-OO; shaded green). For sufficiently thick samples (Bulk, shaded pink), the latter transition is pre-empted by a first-order transition to the bulk state, which, in the Hartree–Fock approximation used here, has ferromagnetic spin order and G-type ($\pi\pi\pi$) antiferro-orbital order. The actual materials exhibit a G-type AF order and a complicated orbital order²⁰ apparently due to subtle lattice distortions neglected here²¹. Inset, schematic of plane perpendicular to (010) direction of (001) superlattice for the $n = 2$ case. Circles show the positions of Sr (white) and La (red) ions respectively; small black dots show the positions of Ti ions.

arguments would predict it to be metallic¹) into a controllable number of planes of SrTiO_3 (a more conventional band-insulating material). Ohtomo *et al.*¹⁴ have measured both the variation of electron density transverse to the planes and the in-plane direct current (d.c.) transport properties of their heterostructure¹⁴. The close chemical similarity of the two component materials minimizes the effects of changes in bandwidths and interaction strengths, allowing us to focus on the consequences of the spreading of electronic charge.

To analyse the experiment of ref. 14 we consider a model consisting of a number n of (001) layers of LaTiO_3 (a Mott insulator with formal Ti valence state d^1) embedded in an infinite SrTiO_3 (a band insulator with formal Ti valence state d^0) crystal, defined theoretically by placing unit charges (corresponding to the difference between the ionic charges of La and Sr) at the ideal crystallographic La positions (see inset to Fig. 1). At very low temperature T , SrTiO_3 is almost ferroelectric: the long-wavelength linear-response dielectric function becomes larger than 10^3 and is sample dependent^{15,16}. The experiments of ref. 14 were performed at room temperature, where the dielectric constant is much smaller, and the quantity of interest to us is the short-length-scale nonlinear response, which is expected to be much smaller. The details of the ferroelectric behaviour are not of interest here, so we model the dielectric properties using a background $\epsilon \approx 15$. We also include

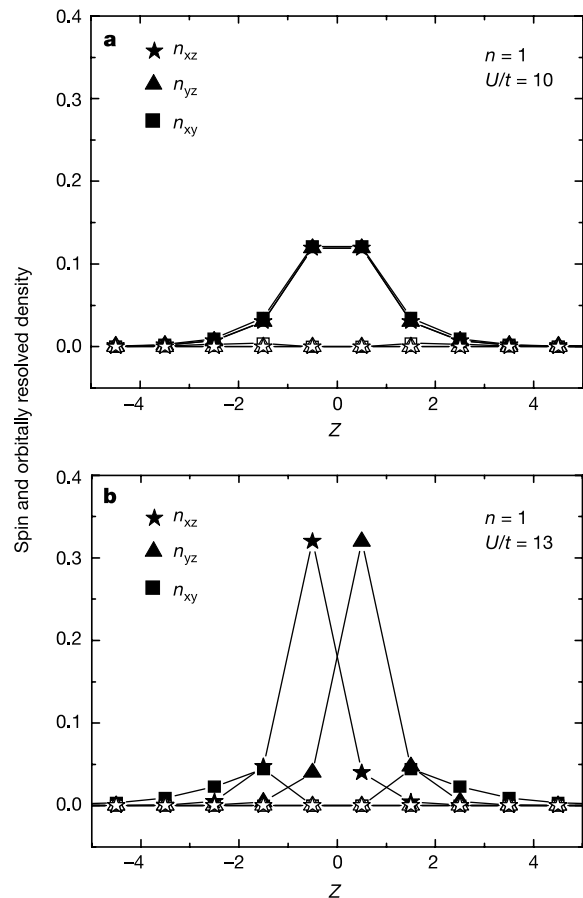


Figure 2 Spin and orbitally resolved charge densities as function of transverse (001) coordinate z for heterostructure with one La layer. La plane is at $z = 0$. Filled (open) symbols indicate majority (minority) spin densities for xz , yz and xy orbitals (see key on figure). **a**, Intermediate U (M-OD) regime; full spin polarization but all three orbitals equally occupied. **b**, Large U (FM-OO) regime. Full spin polarization persists, orbital disproportionation occurs. In the Ti layer at $z = 1/2$, the yz orbital is dominant; in the Ti layer at $z = -1/2$, the xz orbital is dominant; at larger $|z|$ the xy orbital is dominant, but the electron density is low.

(via the standard Hartree approximation) the electric fields produced by the Ti d electrons. Our results depend only weakly on ϵ , for ϵ in the range 5–40, although if the SrTiO₃ actually becomes ferroelectric much more dramatic behaviour might occur. Indeed, D. R. Hamann (personal communication) finds, using a GGA (generalized gradient approximation) band calculation at $T = 0$, that the ferroelectric distortions of the SrTiO₃ overscreen the potential arising from the La ions, so that for the $n = 1$ heterostructure the electrons would become completely detached from the La layers and reside in the centre of the SrTiO₃ layers.

The crucial correlated-electron physics involves the Mott-insulating behaviour of electrons originating from Ti-derived t_{2g} -symmetry orbitals (d_{xy} , d_{xz} and d_{yz}). Within band theory¹⁷, these reside in three interpenetrating quasi-two-dimensional bands (one for each orbital), well-described by a nearest-neighbour tight-binding model of electrons hopping with amplitude t between Ti sites (which we also take to be at their ideal crystallographic positions). The strong correlations responsible for the Mott-insulating behaviour of LaTiO₃ are described by the standard on-site interaction parameters U , U' and J , describing the intraorbital Coulomb, interorbital Coulomb, and interorbital exchange/pair hopping interactions (we use the conventions of ref. 18), respectively. We solve the model via the unrestricted Hartree–Fock approximation, which, although not as accurate for excitation spectra and non-zero temperature properties, gives a reliable picture of the ground-state properties of interest here; especially of the nature of the ordered phases, approximate location of phase boundaries and trends in phase boundaries with changes in parameters¹. Representative parameter values of $t = 0.3$ eV, $\epsilon = 15$, $U' = 7U/9$ and $J = U/9$ (ref. 18) were used. The on-site interaction U was varied, both to explore the range of possible physics and because different measurements^{18,19} imply different interaction strengths. For further specifics of the calculation see <http://www.phys.columbia.edu/millis/oxidehetero.pdf>. Our calculations focus on electronically driven effects and do not include atomic relaxation, which Ohtomo *et al.*¹⁴ asserted was not important in the experimental system. Relaxation (perhaps by amounts below the sensitivity of ref. 14) undoubtedly occurs, but will only enhance the charge-spread effect, see below.

Figure 1 shows our calculated phase diagram in the plane defined

by interaction strength and inverse number of La layers. Varying the quantum well thickness both changes the critical interaction strength required for stabilizing ordered phases and, more importantly, changes the nature of the ordered phase (it causes electronic reconstruction). In particular, bulk behaviour is only obtained for very thick heterostructure. We also note that if the estimates of $U/t \approx 8$ –10 implied by optical conductivity experiments²⁰ are accepted, then our results suggest that an interesting series of phase transitions will occur as the thickness of the layers is varied.

Figure 2 displays specifics of the order occurring in the $n = 1$ heterostructure at different interaction strengths. Figure 2a shows that at $U/t = 10$, the heterostructure exhibits full ferromagnetic spin polarization and no orbital order. (The absence of orbital ordering is seen from the equality of $n_{xz}(z)$ and $n_{yz}(z)$; $n_{xy}(z)$ is allowed by symmetry to differ from n_{xz} and n_{yz} , but for energetic reasons all three densities turn out to be nearly equal at $U/t = 10$.) Figure 2b demonstrates the orbital ordering occurring as U is increased. For intermediate $1 < n \leq 6$ the spin and orbital ordering patterns are similar in plane but alternate from layer to layer. For small n , the only stable states we find are translation-invariant in the xy plane; however, beyond-Hartree–Fock effects arising from the quasi-one-dimensional nature of the xz and yz bands will lead to a very weak instability towards an incommensurate density wave.

We now discuss the physical origin of our results. The dashed curve in Fig. 3 shows the variation of the electron density in the direction transverse to the LaTiO₃ layers for a structure consisting of six LaTiO₃ planes. A ‘leakage’ of electronic charge density out of the LaTiO₃ layers and into the SrTiO₃ region is evident. We have found that the electron density in the SrTiO₃ region does not depend significantly on the value of U or other strong correlation aspects of the problem, but is controlled mainly by the (self-consistently screened) potential arising from the La. In a heterostructure involving different transition-metal ions in the different layers, a difference in on-site potential (electron affinity) would also be important. The transition region also does not change significantly if the ϵ is varied over a reasonable range: relatively robustly, there is an approximately three-unit-cell-wide range over which most of the charge-density drop occurs. In the transition region, interaction effects are weaker owing to the lower (and incommensurate) density, and this, along with the anisotropy implied by the quantum

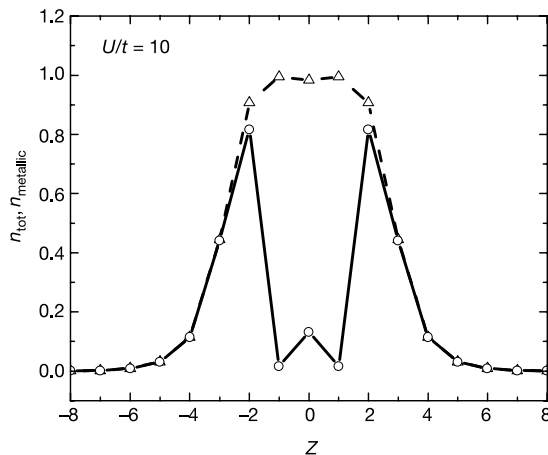


Figure 3 Dependence of total and metallic-subband charge densities $n_{\text{tot}}(z)$, $n_{\text{metallic}}(z)$ on transverse spatial coordinate z for heterostructure with 6 La layers and $U = 10t$. The upper curve (triangles on dashed line) shows electron density as a function of position. An approximately three-unit-cell ‘transition region’ is visible, over which electron density drops from value $n_{\text{tot}}(z) = 1$, characteristic of bulk LaTiO₃, to $n_{\text{tot}}(z) = 0$ characteristic of bulk SrTiO₃. The lower curve (circles on solid line) shows the density of electrons in subbands exhibiting metallic behaviour in the sense defined in the text. The association of metallic behaviour with the near-interface region is evident.

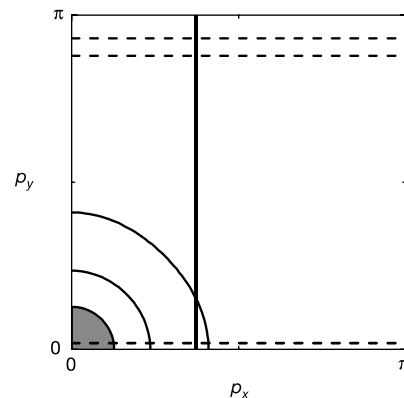


Figure 4 Fermi surface contours as functions of momenta (p_x, p_y) in plane parallel to layers, for six-layer heterostructure, $U = 10t$. The heavy vertical line denotes Fermi surfaces arising from xz -derived subbands, giving most of the charge density at layers $z = \pm 3$ (even–odd splitting $<$ linewidth). The curves are Fermi contours from xy -derived subbands, contributing most of the charge density at layers $z = \pm 4, 5$; the shaded area shows five very slightly occupied xy -derived subbands, accounting for $n_{\text{tot}}(z)$ at $|z| > 5$. The dashed lines are yz -derived bands, contributing most of the charge density at $z = \pm 2$ (nearly full band; even–odd splitting evident) and some of the charge density at $z = \pm 4$ (nearly empty bands; even–odd splitting $<$ linewidth).

well structure, leads to the differences in phase diagram displayed in Fig. 1. Even more importantly, as we now show, the 'transition region' is associated with metallic behaviour.

Adopting semiconductor language, we may say that a set of contiguous LaTiO_3 layers form a quantum well whose electronic states are organized into subbands. The substantial charge spreading means that neither the binding energy nor the inter-subband energy splitting is large, so that in the near-interface region there are several partially occupied subbands, which lead to metallic behaviour. To formulate the problem in a more theoretically precise manner, consider the one-electron spectral function $A(\mathbf{p}, \omega)$ which is the imaginary part of the Fourier transform of the electron expectation value $\langle T\{\psi(\mathbf{r}, t), \psi^\dagger(\mathbf{0}, 0)\} \rangle$. A system is insulating if the chemical potential, μ , lies in a gap, that is, if $A(\mathbf{p}, \omega)$ vanishes for $\omega = \mu$. If not, then A has sharp (quasi-particle) peaks at momenta corresponding to Fermi surface crossings. In the surface or superlattice situations of interest here, the problem has a translation invariance in the plane of the surface or superlattice, so that we may write $A(\mathbf{p}, z, z'; \omega)$, with z the coordinate perpendicular to the plane and $\mathbf{p}$ an in-plane momentum; when $\mathbf{p}$ takes the value $\mathbf{p}_F$ corresponding to an in-plane Fermi surface crossing we have:

$$A(\mathbf{p}, z, z'; \omega) = \pi \sum_a Z^a(\mathbf{p}_F; z, z') \delta(\omega - \mathbf{v}_a \cdot \delta \mathbf{p}_a) + A_{\text{inc}} \quad (1)$$

Here $\mathbf{v}_a$ is the Fermi velocity of subband a , the Z^a values represent the quasiparticle weights at the Fermi surface crossings of the different metallic subbands, the z, z' dependence reflects the spread of the quasiparticle wavefunctions, $\delta \mathbf{p}_a$ labels the distance in momentum space to the relevant Fermi surface crossing, and A_{inc} is the 'incoherent part' of the spectral function, having no sharp structure in momentum space. Metallic behaviour existing near a surface or interface corresponds to $Z^a \neq 0$ for z, z' near the surface or interface.

The solid curve in Fig. 3 displays the charge density associated with the subbands that are metallic in the sense of equation (1) (note that within the Hartree-Fock approximation used here $Z^a = 1$, but that the existence and position of the Fermi surface crossings are reasonably reliably given). The association of metallic behaviour with the near-surface region is evident. Figure 4 shows the Fermi surface contours calculated for the parameters used to construct Fig. 3. The precise positions (and even existence) of the Fermi crossings corresponding to nearly filled or nearly empty bands are unlikely to be a robust feature of our calculation, but the orbital ordering characteristic of intermediate thicknesses, evident from the different locations of the Fermi surface crossings for the xz - and yz -derived bands, along with the general feature of several occupied subbands, is expected to be robust.

Our results are in qualitative agreement with the data of ref. 14, who found charge profiles slightly broader than those shown in Fig. 3, as well as metallic behaviour with effective carrier concentration increasing with n over the range $1 \leq n \leq 6$ studied. We predict that for $n > 6$ the measured carrier concentration will cease to increase linearly with n . We further find that the main contribution to the metallic behaviour arises from quasi-one-dimensional bands. Finally, we comment briefly on the probable effect of lattice relaxation. The main forces in perovskite lattices are coulombic, so to the extent that relaxations occur, they will act to smooth the discontinuity in potential between the LaTiO_3 and SrTiO_3 regions, increasing the charge leakage and therefore the metallicity. For this reason we believe that lattice relaxation effects will not change our results qualitatively, unless the overcompensation effect discussed by D. R. Hamann (personal communication) occurs.

Thus, here we have investigated one 'driver' for novel correlated-electron behaviour at a surface or interface: namely, leakage of charge from one region to another. We have found via a model calculation of an experimentally relevant case¹⁴ that the leakage effect leads to electronic reconstruction, including different forms of magnetic and orbital order different to those found in bulk, and,

generically, metallic behaviour at the interface between the two insulators. □

Received 17 November 2003; accepted 1 March 2004; doi:10.1038/nature02450.

- Imada, M., Fujimori, A. & Tokura, Y. Metal-insulator transitions. *Rev. Mod. Phys.* **70**, 1039–1263 (1998).
- Tokura, Y. & Nagaosa, N. Orbital physics in transition-metal oxides. *Science* **288**, 462–468 (2000).
- Altieri, S., Tjeng, L. H. & Sawatzky, G. A. Ultrathin oxide films on metals: new physics and new chemistry? *Thin Solid Films* **400**, 9–15 (2001).
- Hesper, R., Tjeng, L. H., Heeres, A. & Sawatzky, G. A. Photoemission evidence of electronic stabilization of polar surface in K_3C_{60} . *Phys. Rev. B* **62**, 16046–16055 (2000).
- Duffy, D. M. & Stoneham, A. M. Conductivity and 'negative U' for ionic grain boundaries. *J. Phys. C* **16**, 4087–4092 (1983).
- Altieri, S., Tjeng, L. H. & Sawatzky, G. A. Electronic structure and chemical reactivity of oxide-metal interfaces: $\text{MgO}(100)/\text{Ag}(100)$. *Phys. Rev. B* **61**, 16948–16955 (2000).
- Potthoff, M. & Nolting, W. Metallic surface of a Mott insulator? Mott insulating surface of a metal. *Phys. Rev. B* **60**, 7834–7849 (1999).
- Fang, Z., Solovyev, I. V. & Terakura, K. Phase diagram of tetragonal manganites. *Phys. Rev. Lett.* **84**, 3169–3172 (2000).
- Maiti, K., Mahadevan, P. & Sarma, D. D. Evolution of spectral function in a doped Mott insulator: Surface vs bulk contributions. *Phys. Rev. Lett.* **80**, 2885–2888 (1998).
- Matzendorf, R. *et al.* Ferromagnetism stabilized by lattice distortion at the surface of the p -wave superconductor Sr_2RuO_4 . *Science* **289**, 746–748 (2000).
- Moore, R. G. *et al.* Surface dynamics of the layered ruthenate $\text{Ca}_{1-x}\text{Sr}_x\text{RuO}_4$. *Phys. Status Solidi* (in the press).
- Maiti, K. *et al.* Electronic structure of $\text{Ca}_{1-x}\text{Sr}_x\text{VO}_3$: A tale of two energy scales. *Europhys. Lett.* **55**, 246–252 (2001).
- Izumi, M. *et al.* Perovskite superlattices as tailored materials of correlated electrons. *Mater. Sci. Eng. B* **84**, 53–57 (2001).
- Ohtomo, A., Muller, D. A., Grazul, J. L. & Hwang, H. Y. Artificial charge-modulation in atomic-scale perovskite titanate superlattices. *Nature* **419**, 378–380 (2002).
- Sakudo, T. & Unoki, H. Dielectric properties of SrTiO_3 at low temperatures. *Phys. Rev. Lett.* **26**, 851–853 (1971).
- Müller, K. A. & Burkard, H. SrTiO_3 : an intrinsic quantum paraelectric below 4K. *Phys. Rev. B* **19**, 3593–3602 (1979).
- Fujitani, H. & Asano, S. Full-potential band calculations on YTiO_3 with a distorted perovskite structure. *Phys. Rev. B* **51**, 2098–2102 (1995).
- Mizokawa, T. & Fujimori, A. Unrestricted Hartree-Fock study of transition-metal oxides: Spin and orbital ordering in perovskite-type lattice. *Phys. Rev. B* **51**, 12880–12883 (1995).
- Okimoto, Y., Katsufuji, T., Okada, Y., Arima, T. & Tokura, Y. Optical spectra in $(\text{La}, \text{Y})\text{TiO}_3$: Variation of Mott-Hubbard gap features with change of electron correlation and band filling. *Phys. Rev. B* **51**, 9581–9588 (1995).
- Kiyama, T. & Itoh, M. Presence of $3d$ quadrupole moment in LaTiO_3 studied by $^{47,49}\text{Ti}$ NMR. *Phys. Rev. Lett.* **91**, 167202 (2003).
- Mochizuki, M. & Imada, M. Orbital-spin structure and lattice coupling in RTiO_3 where $R = \text{La}, \text{Pr}, \text{Nd}$, and Sm . *Phys. Rev. Lett.* **91**, 167203 (2003).

Acknowledgements We thank M. Potthoff, G. Sawatzky, W. Ku, H. Y. Hwang, E. W. Plummer and D. R. Hamann for conversations and the US NSF and the JSPS for support.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence should be addressed to A.J.M. (millis@phys.columbia.edu).

Faulting induced by precipitation of water at grain boundaries in hot subducting oceanic crust

Junfeng Zhang^{1,2}, Harry W. Green II¹, Krassimir Bozhilov¹ & Zhenmin Jin²

¹Institute of Geophysics and Planetary Physics and Department of Earth Sciences, University of California, Riverside, California 92521, USA

²Department of Earth Sciences, China University of Geosciences, Wuhan, 430074, China

Dehydration embrittlement has been proposed to explain both intermediate- and deep-focus earthquakes in subduction zones^{1–5}. Because such earthquakes primarily occur at shallow depths or within the core of the subducting plate, dehydration at relatively low temperatures has been emphasized^{6–8}. However, recent careful relocation of subduction-zone earthquakes^{9,10}

shows that at depths of 100–250 km, earthquakes continue in the uppermost part of the slab (probably the former oceanic crust that has been converted to eclogite) where temperatures are higher. Here we show that at such pressures and temperatures, eclogite lacking hydrous phases but with significant hydroxyl incorporated as defects in pyroxene and garnet develops a faulting instability associated with precipitation of water at grain boundaries and the production of very small amounts of melt. This new faulting mechanism satisfactorily explains high-temperature earthquakes in subducting oceanic crust and could potentially be involved in much deeper earthquakes in connection with similar precipitation of water in the mantle transition zone (400–700 km depth). Of potential importance for all proposed high-pressure earthquake mechanisms is the very small amount of fluid required to trigger this instability.

When formed by volcanism at an ocean ridge, the crust is made of basalt, the common black rock of volcanoes like Hawaii and Iceland. However, as it cools and is transported away from the ridge, the rock reacts with sea water circulating through cracks, producing hydrous alteration products. Later, when the crust and associated 50–100 km of underlying mantle are subducted back into the mantle as the return flow of plate tectonics, a series of metamorphic reactions take place that progressively ‘cook’ the water out of the rock, culminating at a depth of approximately 40–70 km in a reaction to form eclogite, a dense rock consisting almost exclusively of garnet and a Na-rich pyroxene, omphacite. Stress concentrations generated by the large volume change accompanying the eclogitization process have been hypothesized to explain intermediate-focus earthquakes within the subducting oceanic crust¹¹; dehydration of amphibolite¹² has also been invoked as a possible trigger of earthquakes in this depth range. At greater depth (>100 km), eclogite may contain small amounts of high-pressure hydrous phases (lawsonite or phengite)¹³ but natural eclogites commonly contain neither of these hydrous phases. Rather, they have either been entirely dehydrated, or they retain hundreds or thousands of parts per million of H₂O bound as OH groups within the nominally anhydrous minerals garnet and pyroxene¹⁴. It is these common eclogites at high temperature and pressure that we address here.

We previously measured the high-temperature rheology of ‘dry’ eclogite (H₂O content of garnet below detection, and pyroxene containing 100–150 p.p.m. H₂O) at a pressure of 3 GPa, and showed it to be ductile under all conditions, exhibiting flow properties similar to those of harzburgite, the type of mantle rock that immediately underlies the oceanic crust¹⁵.

We now have performed similar experiments on ‘wet’ eclogite containing up to 300 p.p.m. H₂O in garnet and 1,000–1,300 p.p.m. H₂O in omphacite, resulting in total rock H₂O content of ~500–800 p.p.m. (Table 1). The experimental material was collected from fresh, coarse-grained, strongly foliated, diamond-facies eclogite at Maobei in the SuLu ‘ultra-high-pressure’ (UHP) metamorphic terrain of China. The rock was disaggregated, cleaned, separated into monomineralic powders, ground to a grain size of 45–75 µm, and reconstituted into eclogite¹⁵. At a strain rate of $1.3 \times 10^{-4} \text{ s}^{-1}$, this rock is ductile at $\leq 1,200 \text{ K}$, with a flow stress in excess of 1,500 MPa (Table 2). However, at 1,300 K and 1,400 K, the material fails by faulting at a stress ($\sigma_1 - \sigma_3$) of ~1,000 MPa, where

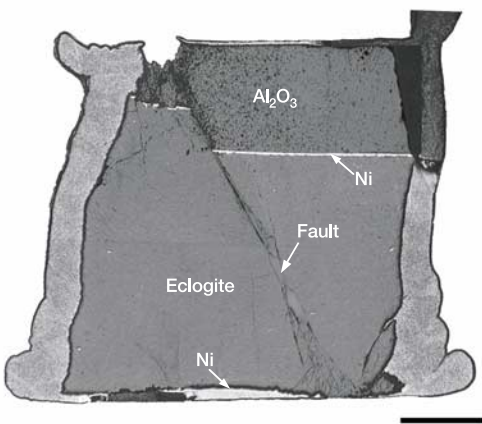


Figure 1 Faulted specimen. Optical micrograph taken in reflected light showing an entire faulted specimen deformed at 3 GPa, 1,400 K and $1.3 \times 10^{-4} \text{ s}^{-1}$ (GB271). Scale bar, 1 mm.

σ_1 is the maximum compressive stress and σ_3 is the confining pressure (Fig. 1). At 1,500 K, the rock is again ductile, with a flow stress less than 600 MPa. Microstructural examination (Fig. 2) shows that specimens deformed at $\geq 1,300 \text{ K}$ also display ultrathin films of brown glass ($\ll 1 \text{ µm}$) at grain boundaries, with progressively more glass at higher deformation temperatures, indicating small degrees of partial melting. Faulted specimens exhibit very large numbers of glass-filled Mode I (tensile) microfractures oriented parallel to σ_1 and glass within the fault zone. Ductile specimens at 1,500 K contained more glass than the faulted specimens but fewer glass-filled Mode I microcracks. Fourier transform infrared spectroscopy of specimens exhibiting evidence of partial melting showed no H₂O remaining within garnet and greatly reduced amounts in omphacite (Table 1). Moreover, specimens showed bubbles within the glass at grain boundaries (Fig. 2d), indicating that either melt and hydrous fluid coexisted during the experiments or the fluid exsolved from the melt during quenching after deformation. The specimens were quenched under stress to below 600 K in less than 6 s, making it unlikely that fluid exsolution occurred during quenching. Moreover, the glass within the fault zone does not contain bubbles (compare Fig. 2c and d). The simplest interpretation of these observations is that melt and fluid separated during failure, implying that precipitation of H₂O to grain boundaries either exceeded the amount needed to saturate the melt or that the kinetics of generation of melt were such that all of the H₂O exsolved did not dissolve into the melt in the time available.

The correlation of faulting with the initiation of partial melting at a temperature 300 K lower than the solidus for the ‘dry’ eclogite studied previously, the finding of fluid bubbles in quenched glass, and the loss of OH from within the minerals of the rock during the experiments, all indicate that H₂O exsolved from both garnet and omphacite beginning at ~1,300 K, accompanied by both melting and faulting. The presence of glass in abundant Mode I microcracks

Table 1 Starting materials and faulted products

Mineral	Percentage	Chemistry	Water concentration in starting materials	Water concentration in faulted products
Garnet	50%	Pyr ₂₃ Alm ₂₀ Gro ₅₇	0–300 p.p.m.	0 p.p.m.
Omphacite	40%	Di ₅₉ Jd ₄₁	1,000–1,300 p.p.m.	600–700 p.p.m.
Kyanite	1%		ND	ND
Rutile	Trace		6,000–8,000 p.p.m.	ND

Water content was measured by a BRUKER A 590 microscope Fourier transform infrared spectrometer with MCT detector at 3,000–4,000 cm⁻¹ spectral range. Both sample and background spectra were collected at 2 cm⁻¹ resolution by an average of 500 scans. ND, not determined. Pyr, pyrope; Alm, almandine; Gro, grossular; Di, diopside; Jd, jadeite

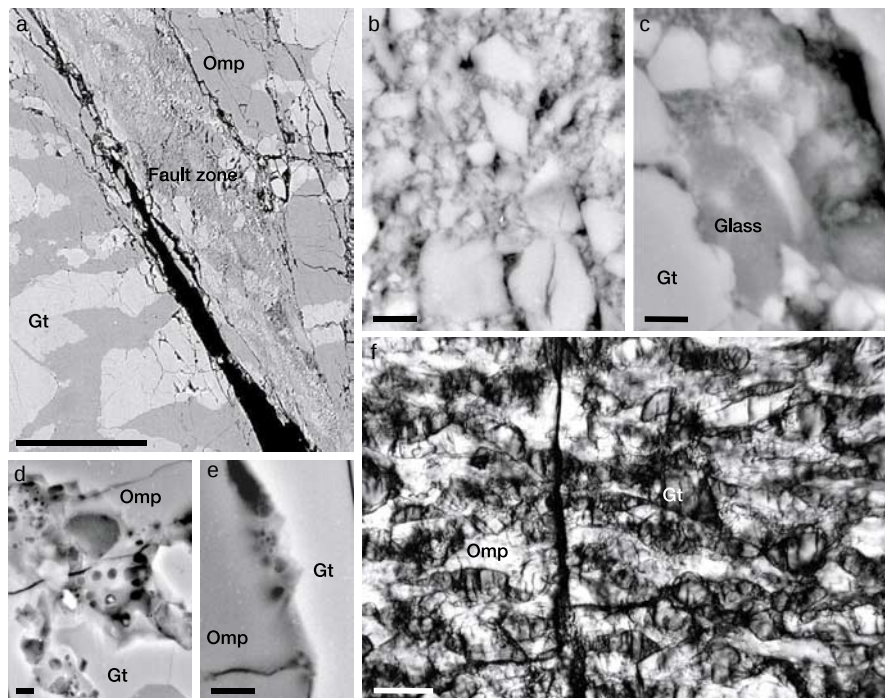


Figure 2 Microstructural details of faulted specimen of Fig. 1. **a–e**, Back-scattered electron SEM images of polished surfaces; **f**, transmission optical image. Fault gouge of variable thickness (**a**) consists of angular fragments with a broad size distribution (**b**), and with glass among the fragments (**c**). Note that the glass lacks vesicles. Away from the fault zones, extremely thin glass films on phase boundaries show abundant vesicles. Panel **d** shows such a grain boundary approximately face-on, whereas **e** shows an oblique view; the vesicles in the glass suggest that $P_{\text{fluid}} = 3 \text{ GPa}$ and that fluid exsolved faster than it

could generate melt (see text). Optical thin section viewed in transmission (**f**) shows all grain boundaries marked with glass films (black) as well as a very large number of glass-filled Mode I microcracks (maximum principal stress top–bottom in image), demonstrating that under stress the melt ($<1\%$) + fluid generated extensive cracking preceding shear failure. Lack of vesicles in fault-zone glass indicates that fluid and melt separated during faulting. Gt, garnet; Omp, omphacite. Scale bars: $50 \mu\text{m}$ (**a**, **f**), $1 \mu\text{m}$ (**b**, **c**, **d**, **e**).

(Fig. 2f) and within fault gouge strongly suggests that failure occurred by pore-pressure-induced reduction of the fracture stress to a value below the flow stress at 1,300 K and 1,400 K. The total amount of melt/fluid generated at these temperatures was $<1 \text{ vol.}\%$, yet it was sufficient to induce faulting. However, at 1,500 K, the melt fraction was 1–2%, significantly reducing the flow stress so that fracture did not occur.

We hypothesized (1) that faulting was induced at 1,300 K and 1,400 K by pore-pressure-induced reduction of the failure stress to below that of the flow stress and (2) that ductile behaviour was restored at 1,500 K because the flow stress was reduced to below that of the fracture stress by an increase in melt fraction. We therefore tested whether an increase in strain rate at 1,500 K could induce faulting, and whether reduction in strain rate at 1,400 K could inhibit faulting. Table 2 shows that at 1,500 K, raising the strain rate by a factor of 3.5 (sample GB275) yielded a higher flow stress but not sufficiently high to reach the fracture stress, whereas raising the strain rate by a factor of 18 (sample GB299) induced faulting. Similarly, reduction of strain rate by a factor of 3 at 1,400 K (sample GB297) reduced the flow stress sufficiently to inhibit fracture.

Thus, we established that under ‘undrained’ conditions (meaning that any fluid generated during the experiment could not escape from the sample), the brittle failure stress for this rock is $975 \pm 15 \text{ MPa}$ and essentially independent of temperature, but perhaps moderately dependent on strain rate over the small range of these parameters tested. Modification of the flow stress by adjusting strain rate within this pressure/temperature window can turn faulting on and off.

We have shown, therefore, that under conditions where the OH defects in omphacite and garnet remain stable, eclogite containing a

few hundred p.p.m. H_2O dissolved in its nominally anhydrous minerals is strong and ductile. But on reaching conditions where hydrous melting can occur, the hydroxyl defects are no longer stable; the dissolved H_2O precipitates uniformly at grain boundaries, and brittle failure occurs under sufficient non-hydrostatic stress. Higher temperatures significantly reduce the flow stress by more extensive melting, hence faulting does not occur; the temperature at which faulting begins and ends can be modulated by adjusting the strain rate.

In the absence of phase changes, the pressure dependence of rock rheology follows the pressure dependence of melting¹⁶, hence these results obtained at 3 GPa can be reasonably extrapolated to any conditions within the stability field of eclogite (2–8 GPa; 60–250 km

Table 2 Experimental conditions and mechanical results

Expt no.	Material	T (K)	Strain rate (s^{-1})	Strength (MPa)	Result	
GB288	Eclogite	1,000	1.3×10^{-4}	$>1,600$	Flowed	Ductile
GB285	Eclogite	1,100	1.3×10^{-4}	$>1,700$	Flowed	
GB284	Eclogite	1,200	1.3×10^{-4}	1,572	Flowed	
GB283	Eclogite	1,300	1.3×10^{-4}	975	Faulted	Brittle
GB274	Eclogite	1,400	1.3×10^{-4}	960	Faulted	
GB271	Eclogite	1,400	1.3×10^{-4}	965	Faulted	
GB304	Eclogite	1,500	1.3×10^{-4}	575	Flowed	Ductile
GB275*	Eclogite	1,500	4.6×10^{-4}	761	Flowed	
GB299*	Eclogite	1,500	2.3×10^{-3}	991	Faulted	
GB297*	Eclogite	1,400	4.6×10^{-5}	833	Flowed	

*These experiments were conducted at various strain rates to turn faulting on and off (see text). All other experiments were conducted at a single strain rate.

depth) for which H_2O exsolves from pyroxene and/or garnet. Solubility can reasonably be inferred to increase with pressure and decrease with temperature on thermodynamic grounds. Experimentally, solubility of H_2O increases with pressure in pyroxene and garnet^{17–19} and appears to decrease with temperature in pyrope garnet¹⁹, but the temperature dependence of solubility is not yet known for pyroxene. At minimum, it is known that H_2O partitions strongly into silicate melts, hence at pressures up to 8 GPa, it is to be expected that crossing of the wet solidus by an eclogite comparable to that investigated here will induce small amounts of hydrous melting at grain boundaries, accompanied by migration into the melt of the H_2O dissolved in pyroxene and garnet, expanding the melt volume and lowering its viscosity. Such dehydration of OH-bearing olivine by incipient melting of peridotite has been predicted²⁰ and confirmed²¹.

On the basis of these results, we propose that under appropriate conditions in eclogitized oceanic crust, H_2O -induced partial melting (or precipitation of small amounts of H_2O incorporated as hydroxyl in nominally anhydrous minerals) could trigger earthquakes in subduction zones at higher temperatures than other proposed faulting mechanisms. Of course, our high-pressure experiments are conducted on specimens that are very small and homogeneous, hence the actual fracture stress to be expected for eclogite in nature would be less than in our experiments, owing to the existence in nature of defects on all scales (including pyroxene-rich zones that may have lower solidi). Indeed, Fig. 2f shows a wide range of sizes of the Mode I precursors of faults in our specimens; in nature, the range can be expected to extend to much greater sizes, leading to a fracture stress that could approximate the frictional sliding stress that would be governed by Byerlee's law, suitably modified for pore pressure effects²².

We emphasize that our specimen material is reconstituted natural eclogite, hence the amount of hydroxyl incorporated into the mineral structures is a natural occurrence, not an experimental construct of dubious relevance to Earth. Indeed, pyroxenes with significantly greater OH contents have been reported from both upper mantle²³ and UHP terranes²⁴. We hypothesize that the H_2O present in our starting materials was retained within the nominally anhydrous phases as pre-existing hydrous minerals were progressively broken down by metamorphism during subduction (also potentially triggering earthquakes¹²). Alternatively, H_2O could be absorbed by eclogite minerals from hydrous fluids generated elsewhere (perhaps by dehydration of serpentine within the mantle portion of the slab) and passing through the subducting oceanic crust at high pressure. Our results also raise the question of whether comparable faulting could be triggered in the mantle transition zone (400–700 km) by exsolution of H_2O from hydrous wadsleyite or ringwoodite^{25,26}.

Finally, we emphasize that the amount of fluid involved in triggering faulting in our experiments is much less than in any previous work, showing that generation of less than 1 vol.% of a low-viscosity fluid phase efficiently delivered to a wide range of sites simultaneously should be able to trigger faulting. Therefore, the two reaction mechanisms previously proposed as triggers for intermediate- and deep-focus earthquakes—dehydration of hydrous phases and transformation-induced faulting^{2,5}—may only require similarly small volumes of the enabling 'fluid'. It follows that hydrous minerals that occur only in low abundance in subducting peridotite (for example, titanian clinohumite²⁷) or small volumes of metastable phases (for example, metastable enstatite at very great depths²⁸) may, if properly distributed, be able to serve as triggers for earthquakes. □

Methods

Specimens were prepared by loading powders into 3-mm-diameter cylindrical Pt capsules with a 25- μm foil of Ni placed above and below the powder. A 1.4-mm-long cylinder of fully dense polycrystalline Al_2O_3 was placed on top of the upper Ni foil within the capsule.

The capsule, approximately 9 mm long, was then welded shut and assembled into the standard high-pressure assembly for our 5-GPa modified Griggs-type deformation apparatus, using CsCl as confining medium. Specimens were pressurized to 3.0 GPa over 15 h at temperature $T = 573$ K, hot-pressed at 1,000–1,300 K for 12 h and then deformed at temperatures of 1,000–1,500 K (Table 2).

Conditions during experimentation were monitored in the following ways: (1) temperature was monitored by two Pt/Rh type B thermocouples included in each experiment, one located near the top of the specimen and the other near the bottom; (2) oxygen fugacity was buffered at Ni/NiO; energy-dispersive (EDS) X-ray analysis of Ni foil revealed the appearance of a small oxygen peak during experimentation, indicating the presence of NiO; (3) potential loss of Fe to the Pt capsule was ruled out by lack of an Fe peak in EDS analysis of Pt after experimentation. No correction was made for the effect of pressure on thermocouples.

Received 19 September 2003; accepted 4 March 2004; doi:10.1038/nature02475.

- Raleigh, C. B. & Paterson, M. S. Experimental deformation of serpentinite and its tectonic implications. *J. Geophys. Res.* **70**, 3965–3985 (1965).
- Green, H. W. & Houston, H. The mechanics of deep earthquakes. *Annu. Rev. Earth Planet. Sci.* **23**, 169–213 (1995).
- Silver, P. G. *et al.* Rupture characteristics of the deep Bolivian earthquake of 1994 and the mechanism of deep-focus earthquakes. *Science* **268**, 69–73 (1995).
- Dobson, D., Meredith, P. G. & Boon, S. A. Simulation of subduction zone seismicity by dehydration of serpentine. *Science* **298**, 1407–1410 (2002).
- Green, H. W. & Marone, C. Instability of deformation. *Rev. Mineral. Geochem.* **51**, 181–199 (2002).
- Peacock, S. M. Are the lower planes of double seismic zones caused by serpentine dehydration in subducting oceanic mantle? *Geology* **29**, 299–302 (2001).
- Omori, S., Kamiya, S., Maruyama, S. & Zhao, D. Morphology of the intraslab seismic zone and devolatilization phase equilibria of the subducting slab peridotite. *Bull. Earthq. Res. Inst.* **76**, 455–478 (2002).
- Hacker, B. R., Peacock, S. M., Abers, G. A. & Holloway, S. D. Subduction factory – 2. Are intermediate-depth earthquakes in subducting slabs linked to metamorphic dehydration reactions? *J. Geophys. Res.* **B 108**, doi:10.1029/2001JB001129 (2003).
- Igarashi, T., Matsuzawa, T., Umino, N. & Hasegawa, A. Spatial distribution of focal mechanisms for interplate and intraplate earthquakes associated with the subducting Pacific Plate beneath the northeastern Japan Arc; a triple-planed deep seismic zone. *J. Geophys. Res.* **106**, 2177–2191 (2001).
- Shelly, D. R. & Beroza, G. C. Downgoing slab seismicity in Japan examined through high precision earthquake hypocenters. *Eos* **83**, F1257 (2002).
- Kirby, S. H., Engdahl, E. R. & Denlinger, R. P. in *Subduction: Top to Bottom* (eds Bebout, G. E. *et al.*) 195–214 (Geophysical Monograph 96, American Geophysical Union, Washington DC, 1996).
- Rushmer, T. in *Subduction: Top to Bottom* (eds Bebout, G. E. *et al.*) 299–306 (Geophysical Monograph 96, American Geophysical Union, Washington DC, 1996).
- Poli, S. & Schmidt, M. W. Petrology of subducted slabs. *Annu. Rev. Earth Planet. Sci.* **30**, 207–235 (2002).
- Rossman, G. Studies of OH in nominally anhydrous minerals. *Phys. Chem. Miner.* **23**, 299–304 (1996).
- Jin, Z.-M., Zhang, J., Green, H. W. II & Jin, S. Eclogite rheology: implications for subducted lithosphere. *Geology* **29**, 667–670 (2001).
- Green, H. W. II & Borch, R. S. The pressure dependence of creep. *Acta Metall.* **35**, 1301–1305 (1987).
- Rauch, M. & Keppler, H. Water solubility in orthopyroxene. *Contrib. Mineral. Petrol.* **143**, 525–536 (2002).
- Ingrin, J. & Skogby, H. Hydrogen in nominally anhydrous upper-mantle minerals: concentration levels and implications. *Eur. J. Mineral.* **12**, 543–570 (2000).
- Lu, R. & Keppler, H. Water solubility in pyrope to 100 kbar. *Contrib. Mineral. Petrol.* **129**, 35–42 (1997).
- Karato, S. Does partial melting reduce the creep strength of the upper mantle? *Nature* **319**, 309–310 (1986).
- Hirth, G. & Kohlstedt, D. L. Water in the oceanic upper mantle: implications for rheology, melt extraction and the evolution of the lithosphere. *Earth Planet. Sci. Lett.* **144**, 93–108 (1996).
- Scholz, C. H. *The Mechanics of Earthquakes and Faulting* 2nd edn, 37–39 (Cambridge Univ. Press, New York, 2002).
- Smyth, J. R., Bell, D. R. & Rossman, R. G. Incorporation of hydroxyl in upper-mantle clinopyroxenes. *Nature* **351**, 732–735 (1991).
- Katayama, I. & Nakashima, S. Hydroxyl in clinopyroxene from the deep subducted crust: evidence for H_2O transport into the mantle. *Am. Mineral.* **88**, 229–234 (2003).
- Smyth, J. R. & Kawamoto, T. Wadsleyite II: A new high pressure hydrous phase in the peridotite- H_2O system. *Earth Planet. Sci. Lett.* **146**, E9–E16 (1997).
- Kohlstedt, D. L., Keppler, H. & Rubie, D. C. Solubility of water in the alpha, beta and gamma phases of $(\text{Mg,Fe})_2\text{SiO}_4$. *Contrib. Mineral. Petrol.* **123**, 345–357 (1996).
- Ulmer, P. & Trommsdorff, V. in *Mantle Petrology: Field Observations and High-pressure Experimentation* (eds Fei, Y., Bertka, C. M. & Mysen, B. O.) 259–281 (Special Publication 6, Geochemical Society, Houston, 1999).
- Hogrefe, A., Rubie, D. C., Sharp, T. G. & Seifert, F. Metastability of enstatite in deep subducting lithosphere. *Nature* **372**, 351–353 (1994).

Acknowledgements We thank L. Dobrzynetska and H. Jung for discussions; F. Forgit for preparation of high-pressure assemblies and apparatus maintenance; and D. Borchardt for advice and assistance with the FTIR spectroscopy. This work was supported by the US National Science Foundation and the Ministry of Science and Technology of China.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to H.W.G. (harry.green@ucr.edu).

Dependence of the duration of geomagnetic polarity reversals on site latitude

Bradford M. Clement

Department of Earth Sciences, Florida International University, SW 8th Street & 107th Avenue, Miami, Florida 33199, USA

An important constraint on the processes governing the geodynamo—the flow in the outer core responsible for generating Earth's magnetic field—is the duration of geomagnetic polarity reversals; that is, how long it takes for Earth's magnetic field to reverse¹. It is generally accepted that Earth's magnetic field strength drops to low levels during polarity reversals, and the field direction progresses through a 180° change while the field is weak¹. The time it takes for this process to happen, however, remains uncertain, with estimates ranging from a few thousand

up to 28,000 years. Here I present an analysis of the available sediment records of the four most recent polarity reversals. These records yield an average estimate of about 7,000 years for the time it takes for the directional change to occur. The variation about this mean duration is not random, but instead varies with site latitude, with shorter durations observed at low-latitude sites, and longer durations observed at mid- to high-latitude sites. Such variation of duration with site latitude is predicted by simple geometrical reversal models, in which non-dipole fields are allowed to persist while the axial dipole decays through zero and then builds in the opposite direction, and provides a constraint on numerical dynamo models.

Recent results based on ⁴⁰Ar/³⁹Ar dating of Brunhes–Matuyama transitional lavas have questioned conventional estimates of reversal duration². These results suggest that the Brunhes–Matuyama reversal lasted 14,000 to 28,000 years instead of a few thousand years—the value that is usually obtained from interpreting sediment and marine magnetic anomaly records. Among the possible reasons for the variation in duration estimates of individual reversals are that: (1) different workers have applied different definitions to define the bounds of reversals; (2) significant uncertainties exist in sedimentation rates or extrusion rates used to determine durations; (3) excursions may have been considered part of the transition zone; (4) the duration of the directional change may vary with site location; or (5) some polarity reversals may have taken longer to occur than others.

The analysis presented here is limited to the past four reversals because records of these reversals are available from a wide geographical range, and there are multiple records of each reversal. The data for this study were taken from the TRANS00 database³ and from records published since the database was updated^{4–6}. Records were selected using the following criteria: (1) each polarity transition record had to exhibit full polarity, nearly antipodal directions both before and after the reversal (falling within the circular standard deviation (CSD) of the full-polarity mean directions); (2) a sedimentation-rate estimate had to be available for the record; and (3) progressive demagnetization had to have been applied. Records

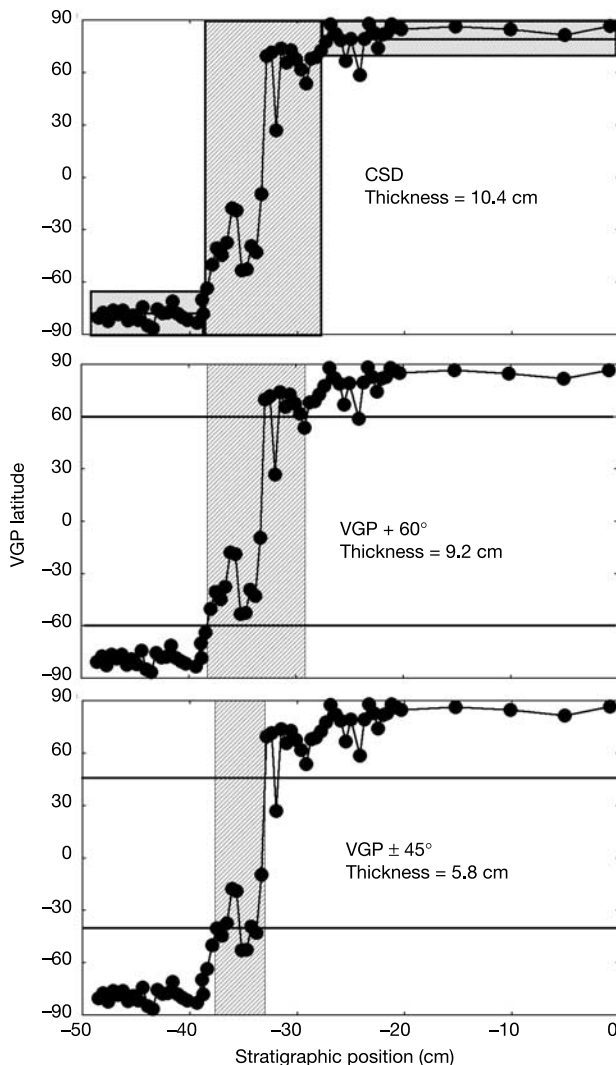


Figure 1 Comparison of the transition-zone thickness in the lower Jaramillo record from ODP Hole 758A (ref. 27) determined using the three methods described in the text. The estimated thickness of the transition zone varies almost by a factor of two, depending on how the transition zone is defined.

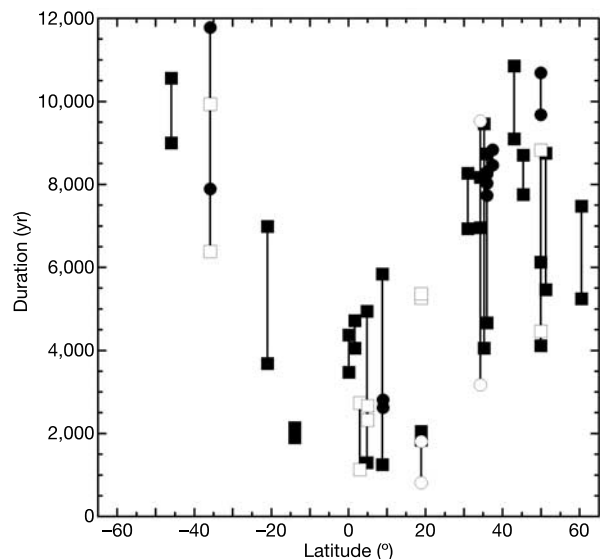


Figure 2 The duration of the directional changes during polarity transitions varies as a function of site latitude. The estimates obtained using the CSD method are taken as the best estimate. However, because this method may in some cases (particularly in low-sedimentation-rate records) overestimate the durations, the estimates obtained using the $VGP \pm 45^\circ$ cut-off method are also shown as lower limits. Results for the Brunhes–Matuyama reversal are indicated by solid squares, the Upper Jaramillo by open circles, the Lower Jaramillo by open squares and the Upper Olduvai by solid circles.

were included where only pilot demagnetization studies were performed in detail, as long as the pilot studies identified a stable endpoint component that is isolated by the selected blanket treatment level. Of the possible records, 30 records meet these criteria and 27 do not.

The thickness of a polarity transition is determined by first defining intermediate-polarity directions, and then by determining their stratigraphic extent. Intermediate-polarity directions may be defined as directions that exceed the typical variation (dispersion) about the full-polarity mean directions. In this approach, a transition zone is defined: the start of the transition zone is the level at which the magnetization directions exceed the CSD⁷ of the mean full-polarity direction. These directions then continue changing until the directions stabilize (Fig. 1) within the CSD of the opposite-polarity mean direction—this level is the end of the transition zone.

In some cases, this method may overestimate the thickness of the directional change. In a low-sedimentation-rate record, the remanence acquisition process may average out the directional changes due to secular variation during the full-polarity interval, resulting in low dispersion about the mean directions (a small CSD). The same smoothing process may increase the thickness of the transition zone by averaging the opposite-, full-polarity directions. Together, these two effects could lead to an overestimate of the transition duration. For this reason, the CSD method provides an upper limit on the estimate of the duration of the directional change.

Intermediate directions have traditionally been identified by selecting an arbitrary angle from the full-polarity directions as a cut-off value. Directions that yield virtual geomagnetic poles (VGPs) that fall more than either 30° or 45° away from the geographical poles have been defined as intermediate. Such a VGP cut-off value introduces a bias towards thinner transition zones at low-latitude sites, because much larger inclination changes at low-latitude sites are required to produce the same shift in VGPs away from the geographical axis than is required at high-latitude sites.

An additional source of potential error arises in either assigning directional swings to the transition zone or treating them as separate excursions. The criterion for determining the boundaries of the

transition zone that the directions must keep changing until they stabilize about the opposite polarity may resolve this uncertainty. But deciding whether the directions actually stabilize is a matter of interpretation. A good relative palaeointensity record (when available) makes it more straightforward to determine whether the rebound is part of the transition, depending on whether or not it occurs during the same geomagnetic low-intensity interval⁸.

Given the possible errors in defining the thickness of the transition intervals, I used both types of methods described above to calculate the transition thickness (Table 1). However, because of the possible latitudinal bias introduced by the VGP cut-off method, the CSD method is considered to provide the best estimate of the transition-zone thicknesses and the transition durations.

Once the transition-zone thicknesses are determined, sedimentation-rate estimates are used to determine the durations of the directional transition. Sedimentation rates in the original studies were calculated using biostratigraphic datums, magnetostratigraphic correlation, and direct or indirect correlation with marine isotope stages. For records published before 1995, sedimentation rates were recalculated using the revised geomagnetic polarity timescale⁹. For several records, the sedimentation rates are not well constrained, and uncertainties in the sedimentation rates are probably the greatest source of error in the duration estimates. The observation that the thicknesses of the transition zones generally increase linearly with increasing sedimentation rate provides confidence that the sedimentation rates are reasonable first approximations (correlation coefficient, $R^2 = 0.9228$, calculated excluding the two very-high-sedimentation-rate Death Valley records¹⁰ that disproportionately control the best-fit line).

The mean of all the durations is $6,992 \text{ yr} \pm 1,089 (2\sigma)$ (Table 1). Intensity changes that are associated with reversals last at least as long and probably longer than the directional changes¹. Therefore, the overall reversal process probably exceeds the maximum average duration of the directional changes observed here (Fig. 2). Unfortunately, enough reliable records of relative palaeointensity through

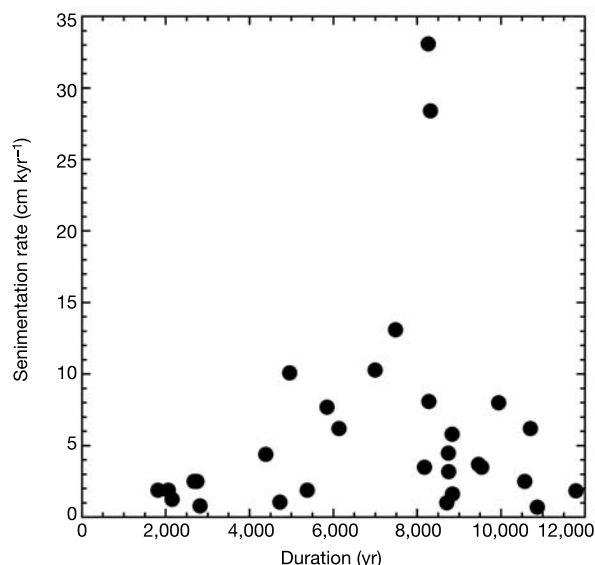


Figure 3 The durations do not correlate with sedimentation rates. This suggests that the observed variation of duration with latitude is not caused by a latitudinal variation in sedimentation rates.

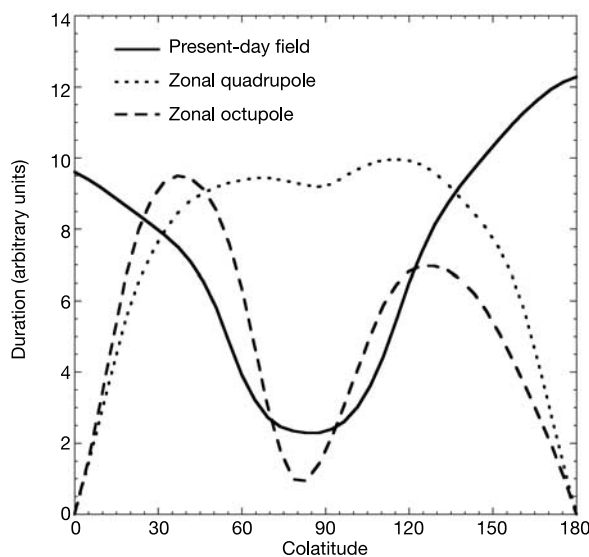


Figure 4 Simple geometrical models in which standing non-dipole fields are allowed to persist while the axial dipole field decays to zero and then builds up in the opposite direction produce transition durations that vary systematically with latitude. The durations for the present-day non-dipole field were calculated for sites distributed along longitude 180°. Durations were calculated in arbitrary time units starting from the beginning of the dipole decrease and continuing to the end of the dipole increase. The octupole and the present-day models present good fits to the observed durations in the mid- to low-latitude range.

transitions are not available to determine a robust estimate of the average duration of the intensity changes.

The variation about the mean duration is not random but instead varies with site latitude (Fig. 2), confirming the observation of this trend made with a much smaller data set almost twenty years ago¹¹. Shorter durations are observed at low-latitude sites, and longer durations are observed at mid- to high-latitude sites. There is no apparent correlation between sedimentation rate and site latitude (Table 1). Nor is there a correlation between sedimentation rate and duration (Fig. 3). Therefore the variation in duration with latitude does not appear to result from a preferential distribution of high-sedimentation-rate records at high latitudes.

The average duration of each reversal does not appear to be significantly different; however, this cannot be assessed until records of the three older reversals are obtained from a wider range of latitudes. The durations of the Brunhes–Matuyama reversal, the lower Jaramillo and the upper Olduvai individual reversals vary with latitude in the same manner as the overall data set. This suggests that the trend is not dependent upon the sense of the reversal (normal-to-reverse or reverse-to-normal polarity). Nor does the trend appear to be unique to a specific reversal. The geographical distribution of the three older reversals, however, together with the smaller number of records of these reversals, do not allow these observations to be tested rigorously.

Variation of duration with site latitude is predicted by simple

reversal models in which non-dipole fields are allowed to persist while the axial dipole decays to zero and then builds up in the opposite direction (Fig. 4). The model based on the present-day field exhibits a variation in duration with latitude similar to that observed. Greater durations occur at latitudes greater than 60°, demonstrating the importance of higher-latitude records for further constraining field geometries. The difference in durations between middle to low latitudes in a reversal dominated by an axial quadrupole is a very small percentage of the total duration of the reversal (less than one time unit). The average difference in the observed durations between mid- to low-latitude sites is 4,000 years. Scaling this difference in the quadrupole model to the observed data, each model time unit would equal approximately 4,000 years, and the modelled reversal would last more than 40,000 years. The octupole model exhibits a greater variation in the duration of the directional change between mid- to low latitudes and provides a better fit with the observed durations (Fig. 4). The difference between the quadrupole and octupole models (Fig. 4) suggests that fields characterized by geometries that are antisymmetric about the Equator may be particularly important during the reversal process.

This modelling effort is not intended to suggest that axially symmetric fields dominated the reversals; the transitional directions in these records are not consistent with purely zonal field geometries. The models discussed here, however, indicate that the variation in the duration of the directional reversal with site latitude

Table 1 Reversal transition records

Reversal	Locality	Ref.	Latitude (°)	Longitude (°)	Thickness (±45 VGP) (cm)	Thickness (CSD) (cm)	Sedimentation rate (cm kyr ⁻¹)	Lower duration (yr)	Upper duration (yr)
Brunhes–Matuyama	V16-58	15	–46.0	30.0	22.50	26.40	2.5	9,000	10,560
	ODP 1082	5	–21.0	11.8	38.00	72.00	10.3	3,689	6,990
	K820305	16	–14.0	174.0	2.40	2.70	1.3	1,905	2,143
	ODP 664	17	0.1	–23.3	15.30	19.26	4.4	3,477	4,377
	RC15-21	18	1.6	–133.0	4.30	5.00	1.1	4,057	4,717
	ODP 767	6	4.8	123.5	13.10	50.00	10.1	1,297	4,950
	ODP 769	6	8.8	121.3	9.63	45.00	7.7	1,250	5,844
	K781030	16	18.9	–160.4	3.50	3.90	1.9	1,842	2,053
	ODP 792A	19	31.0	140.0	56.20	67.00	8.1	6,938	8,272
	Weinan	20	34.2	109.2	24.36	28.60	3.5	6,960	8,171
	Boso Chonan	21	35.2	140.2	15.00	35.00	3.7	4,054	9,459
	Tecopa	22	35.9	–117.0	21.00	39.33	4.5	4,667	8,740
	V20-107	18	43.0	–178.2	6.37	7.60	0.7	9,100	10,857
	V20-108	18	45.4	–179.2	7.91	8.88	1.0	7,755	8,706
	DSDP 609B	23	49.9	–24.2	25.50	38.00	6.2	4,113	6,129
	ODP 883B	24	51.2	167.8	17.50	28.00	3.2	5,469	8,750
	ODP 983	4	60.5	–23.6	68.75	98.00	13.1	5,248	7,481
Mean duration								4,754	6,953
2σ								1,220	1,285
Upper Jaramillo	K781030	25	18.9	–160.4	1.55	3.44	1.9	816	1,811
	Weinan	20	34.2	109.2	11.11	33.36	3.5	3,174	9,531
Mean duration								1,995	5,671
Lower Jaramillo	RC14-14	26	–35.9	60.0	51.11	79.50	8.0	6,389	9,938
	665A	17	2.9	–19.0	2.83	6.84	2.5	1,132	2,736
	ODP 758A	27	5.0	90.0	5.80	6.67	2.5	2,320	2,668
	K781030	25	18.9	–160.4	10.00	10.20	1.9	5,263	5,368
	DSDP 609B	23	49.9	–24.2	25.90	51.20	5.8	4,466	8,828
Mean duration								3,913	5,907
2σ								1,925	3,020
Upper Olduvai	RC14-14	28	–35.9	60.0	14.60	21.80	1.9	7,892	11,784
	K78019	29	9.0	–170.2	2.10	2.25	0.8	2,625	2,813
	Confusion Canyon	10	35.9	–116.6	228.00	236.00	28.4	8,028	8,310
	Death March Canyon	10	35.9	–116.6	256.00	273.30	33.1	7,734	8,257
	K7501	25	37.4	–179.2	13.80	14.40	1.6	8,466	8,834
	DSDP 609	28	49.9	–24.2	60.00	66.30	6.2	9,677	10,694
Mean duration								7,404	8,448
2σ								1,995	2,533
Overall mean duration								4,960	6,992
2σ								988	1,089

Transition thickness determined using the ±45° VGP cut-off and the CSD methods (see text) and the corresponding (lower and upper) durations. Average durations for each reversal are shown. The overall mean represents the average of the durations of all the transition records. 2σ represents two times the standard error of the mean.

may result from straightforward geometrical constraints. This variation may thus be expected in more realistic, numerical simulations of the geodynamo and may provide an important constraint on those models^{12–14}. □

Received 7 November 2003; accepted 2 March 2004; doi:10.1038/nature02459.

- Merrill, R. T. & McFadden, P. L. Geomagnetic polarity transitions. *Rev. Geophys.* **37**, 201–226 (1999).
- Singer, B. S. & Pringle, M. S. Age and duration of the Matuyama-Brunhes geomagnetic polarity reversal from ⁴⁰Ar/³⁹Ar incremental heating analyses of lavas. *Earth Planet. Sci. Lett.* **139**, 47–61 (1996).
- McElhinny, M. W. & Lock, J. IAGA paleomagnetic databases with access. *Surv. Geophys.* **17**, 575–591 (1996).
- Channell, J. E. T. & Lehman, B. The last two geomagnetic polarity reversals recorded in high-deposition-rate sediment drifts. *Nature* **389**, 712–715 (1997).
- Yamazaki, T. & Oda, H. Orbital influence on Earth's magnetic field; 100,000-year periodicity in inclination. *Science* **295**, 2435–2438 (2002).
- Oda, H., Shibuya, H. & Hsu, V. Palaeomagnetic records of the Brunhes/Matuyama polarity transition from ODP Leg 124 (Celebes and Sulu seas). *Geophys. J. Int.* **142**, 319–338 (2000).
- Fisher, R. A. Dispersion on a sphere. *Proc. R. Astron. Soc. A* **217**, 295–305 (1953).
- Channell, J. E. T., Hodell, D. A. & Lehman, B. Relative geomagnetic paleointensity and $\delta^{18}\text{O}$ at ODP Site 983 (Gardar Drift, North Atlantic) since 350 ka. *Earth Planet. Sci. Lett.* **153**, 103–118 (1997).
- Cande, S. C. & Kent, D. V. Revised calibration of the geomagnetic polarity time scale for the Late Cretaceous and Cenozoic. *J. Geophys. Res.* **100**, 6093–6095 (1995).
- Holt, J. W. & Kirschvink, J. L. The upper Olduvai geomagnetic field reversal from Death Valley, California; a fold test of transitional directions. *Earth Planet. Sci. Lett.* **133**, 475–491 (1995).
- Clement, B. M. & Kent, D. V. Latitudinal dependency of geomagnetic polarity transition durations. *Nature* **310**, 488–491 (1984).
- Dormy, E., Valet, J.-P. & Courtillot, V. Numerical models of the geodynamo and observational constraints. *Geochim. Geophys. Geosyst.* **1**, doi:2000GC000062 (2000).
- McMillan, D. G., Constable, C. G., Parker, R. L. & Glatzmaier, G. A. A statistical analysis of magnetic fields from some geodynamo simulations. *Geochim. Geophys. Geosyst.* [online] **28**, doi:2000GC000130 (2001).
- Coe, R. S., Hongre, L. & Glatzmaier, G. A. An examination of simulated geomagnetic reversals from a palaeomagnetic perspective. *Phil. Trans. R. Soc.* **358**, 1141–1170 (2000).
- Clement, B. M. & Kent, D. V. A southern hemisphere record of the Matuyama-Brunhes polarity reversal. *Geophys. Res. Lett.* **18**, 81–84 (1991).
- Theyer, F., Herrero-Bervera, E., Hsu, V. & Hammond, S. R. The zonal harmonic model of polarity transitions; a test using successive reversals. *J. Geophys. Res.* **B 90**, 1963–1982 (1985).
- Valet, J.-P., Tauxe, L. & Clement, B. Equatorial and mid-latitude records of the last geomagnetic reversal from the Atlantic Ocean. *Earth Planet. Sci. Lett.* **94**, 371–384 (1989).
- Clement, B. M., Kent, D. V. & Opdyke, N. D. Brunhes-Matuyama polarity transition in three deep-sea sediment cores. *Phil. Trans. R. Soc. Lond.* **306**, 113–119 (1982).
- Cisowski, S. M. *et al.* Detailed record of the Brunhes/Matuyama polarity reversal in high sedimentation rate marine sediments from the Izu-Bonin Arc. *Proc. ODP Sci. Res.* **126**, 341–352 (1992).
- Zhu, R., Laj, C. & Mazuad, A. The Matuyama-Brunhes and upper Jaramillo transitions recorded in a loess section at Weinan, north-central China. *Earth Planet. Sci. Lett.* **125**, 143–158 (1994).
- Okada, M. & Niitsuma, N. Detailed paleomagnetic records during the Brunhes-Matuyama geomagnetic reversal, and a direct determination of depth lag for magnetization in marine sediments. *Phys. Earth Planet. Inter.* **56**, 133–150 (1989).
- Valet, J.-P., Tauxe, L. & Clark, C. R. The Brunhes-Matuyama transition recorded from Lake Tecopa sediments (California). *Earth Planet. Sci. Lett.* **87**, 463–472 (1988).
- Clement, B. M., Kent, D. V. Geomagnetic polarity transition records from five hydraulic piston core sites in the North Atlantic. In *Init. Rep. Deep Sea Drilling Project* (ed. Ruddiman, W. F. *et al.*) **94**, 831–852 (US Government Printing Office, Washington, 1987).
- Athanassopoulou, J. A *Matuyama-Brunhes Polarity Reversal Record; Comparison Between Thermal and Alternating Field Demagnetization of Ocean Sediments from the North Pacific Transect*. PhD thesis, Univ. California, Santa Barbara, 225 (1993).
- Herrero-Bervera, E. & Theyer, F. Non-axisymmetric behaviour of Olduvai and Jaramillo polarity transitions recorded in north-central Pacific deep-sea sediments. *Nature* **322**, 159–162 (1986).
- Clement, B. M. & Kent, D. V. A detailed record of the lower Jaramillo polarity transition from a southern hemisphere, deep-sea sediment core. *J. Geophys. Res.* **89**, 1049–1058 (1984).
- Gee, J. S. *et al.* Lower Jaramillo polarity transition records from the equatorial Atlantic and Indian oceans. *Proc. ODP Sci. Res.* **121**, 377–391 (1991).
- Clement, B. M. & Kent, D. V. A comparison of two sequential geomagnetic polarity transitions (upper Olduvai and lower Jaramillo) from the Southern Hemisphere. *Phys. Earth Planet. Inter.* **39**, 301–313 (1985).
- Herrero-Bervera, E. & Khan, M. A. Olduvai termination; detailed palaeomagnetic analysis of a north central Pacific core. *Geophys. J. Int.* **108**, 535–545 (1992).

Acknowledgements R. Coe, D. V. Kent, J. Dauphin & B. Midson provided comments that improved the manuscript. J. E. T. Channell and T. Yamazaki provided their data for this work.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to B.M.C. (clementb@fui.edu).

Effectiveness of the global protected area network in representing species diversity

Ana S. L. Rodrigues¹, Sandy J. Andelman³, Mohamed I. Bakarr⁴, Luigi Boitani⁵, Thomas M. Brooks¹, Richard M. Cowling⁶, Lincoln D. C. Fishpool⁷, Gustavo A. B. da Fonseca^{1,8}, Kevin J. Gaston⁹, Michael Hoffmann¹, Janice S. Long², Pablo A. Marquet¹⁰, John D. Pilgrim¹, Robert L. Pressey¹¹, Jan Schipper¹², Wes Sechrest², Simon N. Stuart², Les G. Underhill¹³, Robert W. Waller¹, Matthew E. J. Watts¹⁴ & Xie Yan¹⁵

¹Center for Applied Biodiversity Science, and ²IUCN-SSC/CI-CABS Biodiversity Assessment Unit, Conservation International, Washington, DC 20036, USA

³National Center for Ecological Analysis and Synthesis, University of California, Santa Barbara, California 93101, USA

⁴World Agroforestry Centre (ICRAF), Gigiri Nairobi, Kenya

⁵Dipartimento di Biologia Animale e dell'Uomo, Università di Roma 'La Sapienza', 00185 Rome, Italy

⁶Terrestrial Ecology Research Unit, Department of Botany, University of Port Elizabeth, Port Elizabeth 6000, South Africa

⁷BirdLife International, Cambridge CB3 0NA, UK

⁸Departamento de Zoologia, Universidade Federal de Minas Gerais, Belo Horizonte 31270, Brazil

⁹Biodiversity and Macroecology Group, Department of Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN, UK

¹⁰Center for Advanced Studies in Ecology and Biodiversity (CASEB) and Departamento de Ecología, Facultad de Ciencias Biológicas, Pontificia Universidad Católica de Chile, Casilla 114-D, Santiago, Chile

¹¹New South Wales Department of Environment and Conservation, Armidale, New South Wales 2350, Australia

¹²Department of Fish and Wildlife Resources, University of Idaho, Moscow, Idaho 83844, USA

¹³Avian Demography Unit, Department of Statistical Sciences, University of Cape Town, Rondebosch 7701, South Africa

¹⁴194W Hill Street, Walcha, New South Wales 2354, Australia

¹⁵Institute of Zoology, Chinese Academy of Sciences, Beijing 100080, China

The Fifth World Parks Congress in Durban, South Africa, announced in September 2003 that the global network of protected areas now covers 11.5% of the planet's land surface¹. This surpasses the 10% target proposed a decade earlier, at the Caracas Congress², for 9 out of 14 major terrestrial biomes¹. Such uniform targets based on percentage of area have become deeply embedded into national and international conservation planning³. Although politically expedient, the scientific basis and conservation value of these targets have been questioned^{4,5}. In practice, however, little is known of how to set appropriate targets, or of the extent to which the current global protected area network fulfils its goal of protecting biodiversity. Here, we combine five global data sets on the distribution of species and protected areas to provide the first global gap analysis assessing the effectiveness of protected areas in representing species diversity. We show that the global network is far from complete, and demonstrate the inadequacy of uniform—that is, 'one size fits all'—conservation targets.

Systematic approaches to conservation planning have been developed over the last two decades to guide the efficient allocation of the scarce resources available for protecting biodiversity⁶. Gap analysis is a planning approach based on assessment of the comprehensiveness of existing protected area networks and identification of gaps in coverage^{7,8}. It has also been developed into a formal method now applied by the US Geological Survey National Gap Analysis Program⁹ and others. Numerous gap analyses at regional scales reveal that coverage of biodiversity by existing networks of protected areas is inadequate^{10,11}. Furthermore, many such networks are

Table 1 Numbers of gap species in the current protected area network and in randomly selected networks

Taxon	Median range size (km ²)	Numbers of gap species				
		Current network (all PAs)	Current network (PAs >1,000 ha and IUCN I–IV)	Model I (equal area sites)	Model II (variable area sites)	Model III (tropical bias)
All species						
Mammals (<i>n</i> = 4,735)	247,341	258 (5.5%)	644 (13.5%)	297.7 (6.3%)	342.3 (7.2%)	226.6 (4.8%)
Turtles (<i>n</i> = 273)	309,172	21 (7.7%)	48 (17.6%)	24.6 (9.0%)	26.5 (9.7%)	23.8 (8.7%)
Amphibians (<i>n</i> = 5,454)	7,944	913 (16.7%)	1,718 (31.5%)	1,230.2 (22.6%)	1,507.8 (27.7%)	804.2 (14.7%)
Threatened species						
Mammals (<i>n</i> = 1,063)	22,902	149 (14.0%)	314 (29.6%)	191.8 (18.0%)	218.2 (20.5%)	151.6 (14.3%)
Birds (<i>n</i> = 1,171)	4,015	232 (19.8%)	437 (37.3%)	349.7 (29.9%)	409.6 (35.0%)	275.4 (23.5%)
Turtles (<i>n</i> = 119)	167,611	12 (10.1%)	32 (26.9%)	15.9 (13.3%)	17.3 (14.6%)	15.5 (13.1%)
Amphibians (<i>n</i> = 1,543)	896	411 (26.6%)	767 (49.7%)	604.5 (39.2%)	740.0 (48.0%)	423.3 (27.4%)
All species analysed (<i>n</i> = 11,633)	38,229	1,424 (12.2%)	2,847 (24.5%)	1,902.3 (16.4%)	2,286.2 (19.7%)	1,330.3 (11.4%)

The total numbers of species and their respective median range size are given for comparative purposes. Values in parentheses are the percentage of all species/threatened species analysed within a given taxon. PAs, protected areas.

skewed towards particular ecosystems, often those that are less economically valuable, leaving others inadequately protected¹². At the global scale, however, the degree to which biodiversity is represented within the existing network of protected areas is unknown.

In this analysis, we considered a species to be a 'covered species' if any protected area overlapped any extent of its mapped distribution, and otherwise to be a 'gap species'. Overall, 1,424 gap species (12% of all species analysed) were identified (Table 1). Protected areas may not retain all of their species if they are too small to maintain viable populations¹³ or if they are used extractively¹⁴. Of the covered species, 1,423 were not represented in any protected area larger than 1,000 ha and in stricter conservation classifications (The World Conservation Union (IUCN) categories I–IV¹⁵). Threatened and restricted-range species are those of most conservation concern^{16–18}. Sets of species with smaller median range sizes tend to have a higher proportion of gap species (Table 1). Hence, amphibians are the least represented taxon and, within any given

taxon, threatened species (which tend to have smaller ranges) have proportionally higher numbers of gap species than do all species considered together. Overall, 20% of all threatened species analysed were identified as gap species.

The number of covered species is an overestimate, mainly because of two unrealistic assumptions. First, all protected areas are considered to be adequate for protecting every species, whereas in reality even those classified in IUCN categories I–IV vary substantially in the degree of effectiveness and enforcement¹⁹. Second, it assumes that species can be protected equally effectively in any part of their range, regardless of habitat suitability, and by the protection of any fraction of that range, regardless of viability constraints. In practice, simple presence within a protected area is insufficient to ensure the long-term persistence of many species, particularly those with demanding habitat or area requirements¹³, and does not consider threats such as global climate change²⁰.

As species are only considered to be gap species if they are not touched by any protected area, concentrations of gap species in a

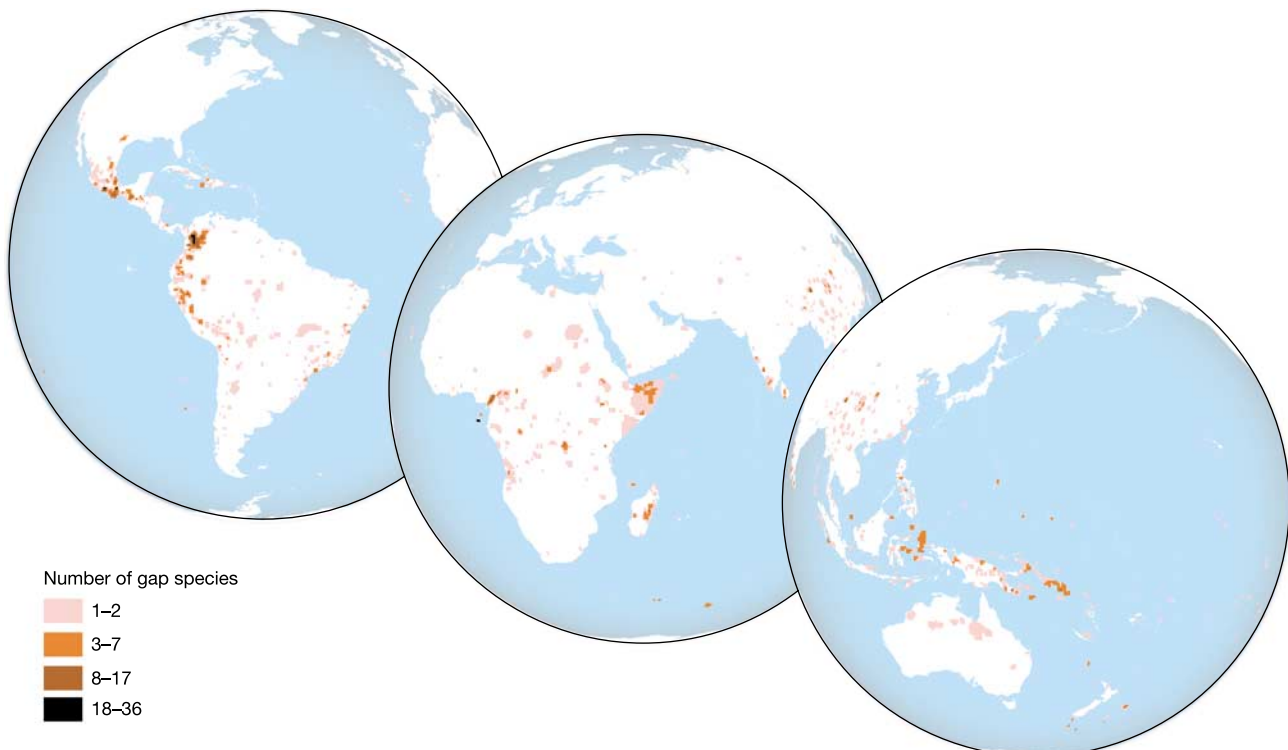


Figure 1 Density map of gap species per half-degree cell, created by overlaying the ranges of all species not covered by any protected area.

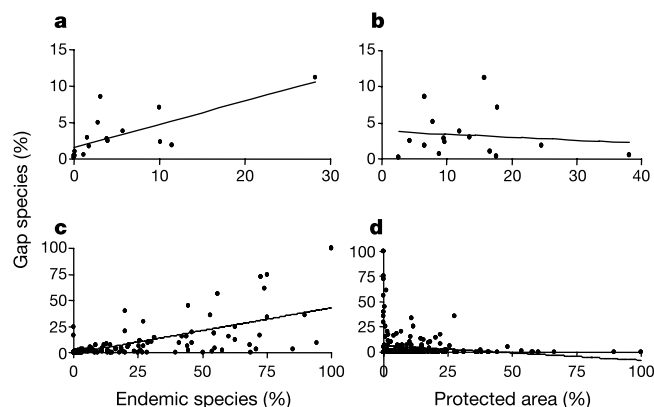


Figure 2 Percentage of gap species in relation to endemism levels and percentage of area protected across biomes and countries. **a–d**, Relationships between: percentage of species in each biome that are endemic and percentage that are gap species (**a**) ($n = 16$, $r = 0.72$, $P < 0.005$); percentage of each biome's protected area and percentage of gap species (**b**) ($P > 0.5$); percentage of species in each country that are endemic and percentage that are gap species (**c**) ($n = 247$, $r = 0.69$, $P < 0.001$); percentage of each country's protected area and percentage of gap species (**d**) ($n = 247$, $r = 0.15$, $P < 0.05$).

given region may be explained by sparse protected area coverage and/or by a concentration of narrowly distributed species. The global distribution of gap species (Fig. 1) is influenced more strongly by the latter. Indeed, within a given biome²¹, the percentage of species that are gaps is highly significantly correlated with the level of endemism, independent of the percentage of area protected (Fig. 2a, b). Across countries, the percentage of gap species decreases with percentage of area protected, but is more strongly correlated with levels of national endemism (Fig. 2c, d). Consequently, although in some regions the absence of protected areas allows for relatively widespread gap species (notably in Somalia), the map of gap species mainly reflects the presence of narrowly distributed species (Fig. 1). The regions highlighted include many widely recognized centres of endemism^{16,22}, such as Yunnan province and the mountains surrounding the Sichuan basin in southern China, the Western Ghats of India, Sri Lanka, the islands of Southeast Asia and Melanesia, the Pacific islands, Madagascar, the Cameroon highlands, Mesoamerica, the tropical Andes, the Caribbean, and the Atlantic Forest of South America. Most of these are montane or insular regions in the tropics.

These results have implications for global conservation planning strategies, as they clearly demonstrate that the percentage of area already protected in a given country or biome is a very poor indicator of additional conservation needs. Contrary to frequent recommendations^{1,23}, current protection levels should not be used as a significant criterion to guide priorities for allocation of future conservation investments. Indeed, the regions with greatest need for expansion of the global protected area network are not necessarily those with a lower percentage of their area protected; rather, they typically are those with higher levels of endemism²⁴. Conversely, uniform targets based on percentage of area protected (except for 100%) cannot be used as a ceiling to distinguish between regions sufficiently protected and those that need additional protection^{4–5}.

Global conservation strategies based on the recommendation that 10% (or other similar targets) of each country or biome be protected will not be effective because they are blind to the fact that biodiversity is not evenly distributed across the planet²⁵; by the same token, neither should protected areas be. Indeed, a network with the same total area as the existing one but evenly distributed across the world would perform less adequately than the current network in

representing species of mammals, amphibians, turtles and threatened birds (Table 1). The better performance of the current network indicates uneven distribution of protected areas relative to biodiversity pattern. Indeed, the current network is significantly (albeit not overwhelmingly) biased towards sites with higher richness of all species, restricted-range species and threatened species. This may be the legacy of decisions to locate some protected areas in better sites, and/or be symptomatic of higher levels of biodiversity loss outside protected areas¹⁹. Nonetheless, the current global network could still perform better in terms of species coverage. For example, a network biased towards the tropics (to match their higher level of endemism) would have fewer gap species than the current network, and far fewer gap species than a random unbiased network (Table 1).

Our results demonstrate that if the conservation goal is species representation, then the expansion of the global network of protected areas must account for biodiversity patterns, rather than rely on general percentage-based targets that are formed largely by political and feasibility considerations^{4–5}. Given the increasing threats to biodiversity, such expansion should be made strategically by focusing on those regions that would contribute most to the global system and prioritizing, within those, the regions where the urgency for conservation action is greatest²². Conservation strategies must also address the complexity of natural ecosystems, including genetic and phylogenetic diversity, and ecological and evolutionary processes²⁶.

The existing protected area network provides an invaluable service in shielding habitat from destructive use and hence in reducing biodiversity loss¹⁹. However, our global gap analysis clearly demonstrates that the global protected area network is still far from complete, even for terrestrial vertebrates, the best known and most popular of all species groups²⁷. Of the species considered, at least 12% are not represented in any protected area, despite the extremely strict assumptions applied for identifying gap species. It is likely that other taxa with high levels of endemism, such as plants and insects, are even less well represented, given the tendency for sets of species with smaller range sizes to have higher proportions of gap species.

Protected areas are not the only tactic available to conservation planners, but they are highly cost effective in protecting biodiversity²⁸. Advances in data availability and in the science of conservation planning enable us to act strategically in the face of increasing human pressure. Clearly, the task ahead is as urgent as it is challenging, as much biodiversity remains to be protected. □

Methods

Data

Data on the global distribution of protected areas were obtained from the 2003 World Database on Protected Areas²⁹. Distribution maps were obtained for 11,633 species of terrestrial vertebrates: 4,735 terrestrial mammals, compiled by the IUCN Global Mammal Assessment; 1,171 globally threatened birds¹⁷; 273 freshwater turtles and tortoises³⁰; and 5,454 amphibians, compiled by the IUCN Global Amphibian Assessment. These species data also include assessments of conservation status, with 1,063 mammals, 1,171 birds, 119 turtles and 1,543 amphibians having been listed as globally threatened by the IUCN Red List^{17–18}. See Supplementary Information for more details.

Randomly distributed networks

Two null models were created to simulate a network of protected areas with similar characteristics to the existing one, but evenly spread around the world: Model I (equal area sites), 69,794 circles, of the same size as the mean area of a protected site, and 11,119 points were randomly spread around the world's land surface (excluding Antarctica); Model II (variable area sites), 69,794 circles, with the same distribution of sizes as the current protected area network, and 11,119 points were randomly spread around the world's land surface (excluding Antarctica).

Of all species that are restricted to either the tropical or the non-tropical regions (that is, excluding species that span both), 75.8% are found in the tropics; however, only 45.8% of the global protected area network is in the tropics. Therefore, we considered a third model in which the percentage of the global protected area in the tropics was increased to match its level of endemism: Model III (tropical bias), 69,794 circles, of the same size as the mean area of a protected site, and 11,119 points were distributed such that 75.8% of each occurred in the tropics, having random distributions within tropical and non-tropical areas.

Sixty replicates were obtained for each of these randomly distributed networks. These were then overlaid with species distributional data to analyse the number of gap species in each case. See Supplementary Information for the confidence intervals for each of the models.

Richness of protected and unprotected cells

The richness of each quarter-degree cell touching land (outside Antarctica) was calculated for all species, restricted-range species¹⁶ (occupying $\leq 50,000 \text{ km}^2$) and threatened species. Cells touching protected areas were considered 'protected'. Protected cells are significantly ($P < 0.001$) biased towards higher richness of all, restricted-range and threatened species. See Supplementary Information for a comparison of frequency distributions.

Received 21 December 2003; accepted 11 February 2004; doi:10.1038/nature02422.

- Chape, S., Fish, L., Fox, P. & Spalding, M. *United Nations List of Protected Areas* (IUCN/UNEP, Gland, Switzerland/Cambridge, UK, 2003).
- The World Conservation Union. *Parks For Life: Report of the IVth World Congress on National Parks and Protected Areas* (IUCN, Gland, Switzerland, 1993).
- Kamden-Toham, A. *et al.* Forest conservation in the Congo Basin. *Science* **299**, 346 (2003).
- Soulé, M. E. & Sanjayan, M. A. Conservation targets: do they help? *Science* **279**, 2060–2061 (1998).
- Pressey, R. L., Cowling, R. M. & Rouget, M. Formulating conservation targets for biodiversity pattern and process in the Cape Floristic Region, South Africa. *Biol. Conserv.* **112**, 99–127 (2003).
- Margules, C. R. & Pressey, R. L. Systematic conservation planning. *Nature* **405**, 243–253 (2000).
- Scott, J. M. *et al.* Gap analysis—a geographic approach to protection of biological diversity. *Wildl. Monogr.* **123**, 1–41 (1993).
- Lacher, T. E. Jr in *GIS Methodologies for Developing Conservation Strategies* (eds Savitsky, B. G. & Lacher, T. E. Jr) 199–209 (Columbia Univ. Press, New York, 1998).
- Jennings, M. D. Gap analysis: Concepts, methods, and recent results. *Landscape Ecol.* **15**, 5–20 (2000).
- Scott, J. M. *et al.* Nature reserves: Do they capture the full range of America's biological diversity? *Ecol. Appl.* **11**, 999–1007 (2001).
- Andelman, S. J. & Willig, M. R. Present patterns and future prospects for biodiversity in the Western Hemisphere. *Ecol. Lett.* **6**, 818–824 (2003).
- Pressey, R. L. Ad hoc reservations—Forward or backward steps in developing representative reserve systems? *Conserv. Biol.* **8**, 662–668 (1994).
- Newmark, W. D. Insularization of Tanzanian parks and the local extinction of large mammals. *Conserv. Biol.* **10**, 1549–1556 (1996).
- Peres, C. A. & Lake, I. R. Extent of nontimber resource extraction in tropical forests: accessibility to game vertebrates by hunters in the Amazon basin. *Conserv. Biol.* **17**, 521–535 (2003).
- The World Conservation Union. *Guidelines for Protected Area Management Categories* (IUCN CNPPA/WCMC, Gland, Switzerland/Cambridge, UK, 1994).
- Stattersfield, A. J., Crosby, M. J., Long, A. J. & Wege, D. C. *Endemic Bird Areas of the World—Priorities for Biodiversity Conservation* (BirdLife International, Cambridge, UK, 1998).
- BirdLife International. *Threatened Birds of the World* (Lynx Edicions/BirdLife International, Barcelona/Cambridge, UK, 2000).
- The World Conservation Union. *IUCN Red List of Threatened Species* [online] (<http://www.redlist.org>) (2003).
- Bruner, A. G., Gullison, R. E., Rice, R. E. & Fonseca, G. A. B. Effectiveness of parks in protecting tropical biodiversity. *Science* **291**, 125–128 (2001).
- Thomas, C. D. *et al.* Extinction risk from climate change. *Nature* **427**, 145–148 (2004).
- Olson, D. M. *et al.* Terrestrial ecoregions of the world: A new map of life on earth. *Bioscience* **51**, 933–938 (2001).
- Myers, N., Mittermeier, R. A., Mittermeier, C. G., Fonseca, G. A. B. & Kent, J. Biodiversity hotspots for conservation priorities. *Nature* **403**, 853–858 (2000).
- Green, M. J. B. & Paine, J. *State of the World's Protected Areas at the End of the Twentieth Century* (WCMC, Cambridge, UK, 1997).
- Rodrigues, A. S. L. & Gaston, K. J. How large do reserve networks need to be? *Ecol. Lett.* **4**, 602–609 (2001).
- Gaston, K. J. Global patterns in biodiversity. *Nature* **405**, 220–227 (2000).
- Cowling, R. M. & Pressey, R. L. Rapid plant diversification: planning for an evolutionary future. *Proc. Natl Acad. Sci. USA* **98**, 5452–5457 (2001).
- Gaston, K. J. & May, R. M. Taxonomy of taxonomists. *Nature* **356**, 281–282 (1992).
- Balmford, A. *et al.* Economic reasons for conserving wild nature. *Science* **297**, 950–953 (2002).
- World Database on Protected Areas. *World Database on Protected Areas* (IUCN-WCPA/UNEP-WCMC, Washington DC, 2003).
- Iverson, J. B., Kiestler, A. R., Hughes, L. E. & Kimerling, A. J. *The EMyS System World Turtle Database* 2003 [online] (<http://emys.geo.orst.edu>) (2003).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank the Moore Family Foundation, the Howard Gilman Foundation and the National Center for Ecological Analysis and Synthesis of the University of California Santa Barbara for support. The analysis was possible thanks to the combined effort of the thousands of individuals and hundreds of institutions who collected and compiled the data, or provided financial support for such efforts. We are grateful to the numerous individuals who contributed to this analysis, especially to K. Buhlmann, S. Butchart, N. Cox, P. P. van Dijk, J. Iverson, R. Kiestler, T. Lacher and B. Young. H. Possingham made valuable comments on the manuscript. Figure 1 was generated by J. Seeber.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.S.L.R. (a.rodriques@conservation.org).

Spatial structure often inhibits the evolution of cooperation in the snowdrift game

Christoph Hauert & Michael Doebeli

Departments of Zoology and Mathematics, University of British Columbia, 6270 University Boulevard, Vancouver, British Columbia V6T 1Z4, Canada

Understanding the emergence of cooperation is a fundamental problem in evolutionary biology¹. Evolutionary game theory^{2,3} has become a powerful framework with which to investigate this problem. Two simple games have attracted most attention in theoretical and experimental studies: the Prisoner's Dilemma⁴ and the snowdrift game (also known as the hawk–dove or chicken game)⁵. In the Prisoner's Dilemma, the non-cooperative state is evolutionarily stable, which has inspired numerous investigations of suitable extensions that enable cooperative behaviour to persist. In particular, on the basis of spatial extensions of the Prisoner's Dilemma, it is widely accepted that spatial structure promotes the evolution of cooperation^{6–8}. Here we show that no such general predictions can be made for the effects of spatial structure in the snowdrift game. In unstructured snowdrift games, intermediate levels of cooperation persist. Unexpectedly, spatial structure reduces the proportion of cooperators for a wide range of parameters. In particular, spatial structure eliminates cooperation if the cost-to-benefit ratio of cooperation is high. Our results caution against the common belief that spatial structure is necessarily beneficial for cooperative behaviour.

The Prisoner's Dilemma illustrates that cooperating individuals are prone to exploitation, and that natural selection should favour cheaters. In this game, two players simultaneously decide whether to cooperate or defect. Cooperation results in a benefit b to the recipient but incurs a cost c to the donor ($b > c > 0$). Mutual cooperation thus pays a net benefit of $R = b - c$, whereas mutual defection results in payoff $P = 0$ for both players. With unilateral cooperation, defection yields the highest payoff, $T = b$, at the expense of the cooperator bearing the cost $S = -c$. It follows that it is best to defect regardless of the co-player's decision. Thus, defection is the evolutionarily stable strategy, even though all individuals would be better off if they all cooperated. This outcome is a simple consequence of the ranking of the four payoff values: $T > R > P > S$. Despite this seemingly convincing argument, many natural species show altruism, with individuals bearing costs to the benefit of others: vampire bats share blood⁹, alarm calls warn from predators¹⁰, monkeys groom each other¹¹, and fish inspect predators preferentially in pairs¹².

In field and experimental studies it is often difficult to assess the fitness payoffs for different behavioural patterns, and even the proper ranking of the payoffs is challenging^{13,14}. This has led to a considerable gap between theory and experimental evidence, and to an increasing discomfort with the Prisoner's Dilemma as the only model to discuss cooperative behaviour^{15,16}. The snowdrift game is a viable and biologically interesting alternative. It differs from the Prisoner's Dilemma in that the payoffs P and S have a reverse order: $T > R > S > P$. This changes the situation fundamentally and leads to persistence of cooperation.

To illustrate the snowdrift game, imagine two drivers that are caught in a blizzard and trapped on either side of a snowdrift. They can either get out and start shovelling (cooperate) or remain in the car (defect). If both cooperate, they have the benefit b of getting home while sharing the labour c . Thus, $R = b - c/2$. If both defect, they do not get anywhere and $P = 0$. If only one shovels, however, they both get home but the defector avoids the labour cost and gets

$T = b$, whereas the cooperator gets $S = b - c$. If costs are high ($2b > c > b > 0$), these payoffs recover the Prisoner's Dilemma. By contrast, if $b > c > 0$, the payoffs generate the snowdrift game, in which the best action depends on the co-player: to defect if the other cooperates, but to cooperate if the other defects. This leads to stable coexistence of cooperators and defectors in well-mixed populations.

According to the replicator dynamics¹⁷, the equilibrium frequency of cooperators in the snowdrift game is $1 - r$, where $r = c/(2b - c)$ is the cost-to-benefit ratio of mutual cooperation. Note, however, that the average population payoff at evolutionary equilibrium is smaller than the average payoff in a population of only cooperators, as in the Prisoner's Dilemma. Thus, the paradox of cooperation is also apparent in the snowdrift game.

An important insight is that spatial structure can promote persistence of cooperation. In particular, if the Prisoner's Dilemma is played in spatially structured populations, in which individuals interact only within a limited local neighbourhood, then cooperation can be maintained. Here we investigate the effects of spatial structure in the snowdrift game. To model spatial structure, we assume that individuals occupy sites on a regular lattice. Whenever a site is updated, the present occupant and its nearest neighbours compete to populate the site with their offspring. Generalizing the replicator dynamics to lattices, competitive success is determined according to differences between the payoffs that each potential parent obtained from game interactions with their nearest neighbours. Updating can be synchronous across the lattice, describing populations with discrete, non-overlapping generations, or asynchronous, describing populations with overlapping genera-

tions in continuous time (see Methods).

Intriguingly, spatial structure fails to enhance cooperation in the snowdrift game and actually tends to reduce the proportion of cooperators. Figure 1 shows equilibrium proportions of cooperators in spatial populations as a function of the cost-to-benefit ratio $r = c/(2b - c)$. Only for small r (high benefits, low costs) is the proportion of cooperators higher than the $1 - r$ expected in well-mixed populations (Fig. 1, dotted line). By contrast, spatial structure favours defectors for larger r . The threshold above which the proportion of defectors is higher than in well-mixed populations depends on the lattice geometry (Fig. 1) and decreases with increasing neighbourhood size N . In all cases, cooperation is eliminated altogether for sufficiently high r , which is again in stark contrast to the well-mixed case. Ultimately, these results are due to the small interaction neighbourhoods in spatially structured populations. Although the qualitative results do not depend on the exact number of neighbours N , some quantitative features do, such as the extinction thresholds for cooperators and defectors (Fig. 1). For example, cooperators vanish near the r value for which a single cooperator in a given neighbourhood constitutes a higher frequency of cooperation than the well-mixed expectation, that is, for which $1/N > 1 - r$.

For an intuitive understanding of the contrary effects of spatial structure in the Prisoner's Dilemma and in the snowdrift game, it is useful to look at snapshots of spatial configurations at stochastic equilibrium near the extinction threshold of cooperators (see Virtual Labs in evolutionary game theory: <http://www.univie.ac.at/virtuallabs>). In the spatial Prisoner's Dilemma, cooperators can survive by forming large, compact clusters (Fig. 2a), thus reducing exploitation by defectors. By contrast, in the spatial snowdrift game cooperators form small filament-like clusters (Fig. 2b). These spatial patterns arise from microscopic processes that are dictated by the payoff structure of the snowdrift game, which makes it advantageous to adopt strategies that are opposite to neighbouring strategies. As a consequence, an isolated cooperator acts as a seed for expanding dendritic structures, but lacks the ability to give rise to compact clusters (Fig. 2c). On average, these emergent spatial

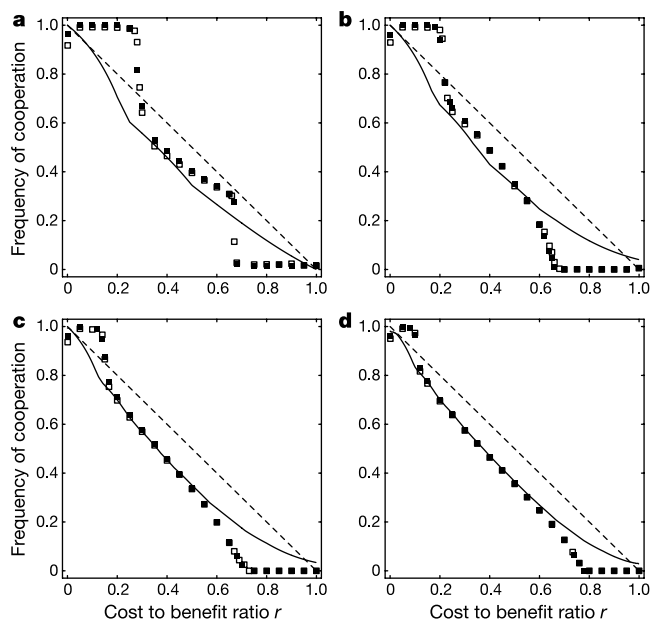


Figure 1 Frequency of cooperators as a function of the cost-to-benefit ratio $r = c/(2b - c)$ in the snowdrift game for different lattice geometries. **a**, Triangular lattice, neighbourhood size $N = 3$; **b**, square lattice, $N = 4$; **c**, hexagonal lattice, $N = 6$; **d**, square lattice, $N = 8$. For small r , spatial structure promotes cooperation; however, for intermediate and high r , the fraction of cooperators is lower than in well-mixed populations (dotted line). This result is largely independent of whether updating is synchronous (filled squares) or asynchronous (open squares). The tendency is correctly predicted by pair approximations (unbroken line), but pair approximation underestimates the effects of local configurations at high and low r . In individual-based simulations, the range of coexistence of cooperators and defectors is delimited by two threshold values: below r_1 defectors vanish, whereas above r_2 cooperators are doomed. Both thresholds correlate with the fate of local configurations: near r_1 defector pairs tend to annihilate and vanish, whereas near r_2 single cooperators and cooperator pairs cannot survive in a sea of defectors. See Methods for simulation details.

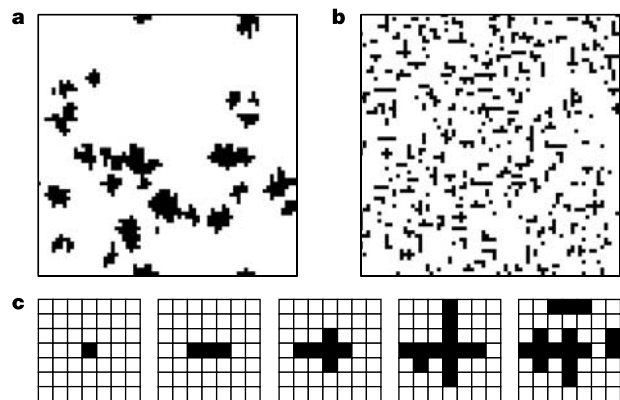


Figure 2 Snapshots of equilibrium configurations of cooperators (black) and defectors (white) in the spatial Prisoner's Dilemma and spatial snowdrift game on a square lattice with $N = 4$ neighbours near the extinction threshold of cooperators. **a**, In the Prisoner's Dilemma, cooperators survive by forming compact clusters ($R = 1$, $T = 1.07$, $S = -0.07$, $P = 0$). **b**, In the corresponding snowdrift game, cooperators are spread out, forming many small and isolated patches ($r = 0.62$; that is, $R = 1$, $T = 1.62$, $S = 0.38$, $P = 0$). This result also holds for other lattice structures (not shown). **c**, Microscopic pattern formation in the spatial snowdrift game. An isolated cooperator can grow into a row of cooperators and then form cross-like structures; however, cooperators cannot expand to compact clusters because the payoff structure protects the defectors in the corners. Eventually, cooperators form a dendritic skeleton. Occasionally, dendrites break off to form new seeds.

patterns generate an advantage for defectors, owing to increased exploitation in the fractal-like zone of contact between the two strategies. This leads to an overall reduction in cooperators as compared with well-mixed populations.

The relevant spatial pattern formation is most evident near the extinction threshold of cooperators, that is, in the parameter region in which the detrimental effects of spatial structure on cooperation are most pronounced. For lower values of the cost-to-benefit ratio r , additional mechanisms that favour cooperation start to be important. For example, despite the fact that it is always better to adopt strategies that are opposite to neighbouring strategies, it is clearly advantageous to have cooperating neighbours. Accordingly, clustering of cooperators can lead to the extinction of defectors for low r , which is in line with the established view that spatial structure should benefit cooperation. We have confirmed the results from our individual-based models using the technique of pair approximation (see Methods and Supplementary Information). Results from this deterministic approximation of the spatial dynamics (Fig. 1, unbroken line) are in good agreement with the stochastic simulations. We also note that our main findings are robust with respect to variations in lattice geometry, in synchrony of updating and in update rules (see Supplementary Information).

When studying cooperation, it is often useful to allow for continuously varying degrees of cooperative behaviour, which can be achieved by considering mixed strategies describing an individual's propensity to cooperate. This approach has been often applied to the hawk–dove game^{3,18}, a version of the snowdrift game traditionally used in behavioural ecology: when competing for resources or mates, hawks escalate conflicts, whereas doves are conciliatory. When two doves meet they share the resource β and

both get $R = \beta/2$, but when facing an escalating hawk the dove takes flight ($S = 0$) and the hawk gets the whole resource ($T = \beta$). If two hawks meet, they escalate until one is injured and incurs a fitness loss γ ($\gamma > \beta$). Escalation thus yields, on average, $P = (\beta - \gamma)/2 < 0$. With $b = (\beta + \gamma)/2$ and $c = \beta$, this game is equivalent to the snowdrift game in the sense that the payoff matrices only differ by a constant, so that update rules based on payoff differences yield identical results. In particular, for the replicator dynamics in well-mixed populations, the evolutionarily stable mixture consists of r hawks and $1 - r$ doves, where $r = \beta/\gamma$. This game can be viewed as a mixed strategy game, in which one individual adopts the behavioural patterns of hawk and dove with specific probabilities¹⁹. In correspondence with the pure strategy game, the evolutionarily stable mixed strategy plays dove with a probability $1 - r$.

We have investigated the mixed-strategy hawk–dove game in spatially structured populations (see Methods). For synchronously updated populations (Fig. 3, filled squares), spatial structure systematically lowers the probability to show dove-like behaviour. In particular, cooperation vanishes for high r , for which all conflicts escalate. For $N = 3$ this happens for $r \geq 0.8$ (Fig. 3a), whereas the corresponding well-mixed populations (Fig. 3a, dotted line) support up to 20% dove-like behaviour. This effect decreases for increasing N , such that results for $N = 8$ (Fig. 3d) are essentially indistinguishable from well-mixed populations. With asynchronous updating (Fig. 3, open squares), the effects of spatial structure are negligible, independent of the lattice geometry and neighbourhood size.

Our results show that spatial extension generally fails to promote cooperative behaviour in the hawk–dove or snowdrift game. In fact, with the exception of small cost-to-benefit ratios, spatial structure tends to reduce the level of cooperation. In contrast to the Prisoner's Dilemma, the snowdrift game is a simple model for the evolution of cooperation when defection is not an evolutionarily stable strategy. We therefore conclude that spatial structure may be rarely beneficial, but often detrimental, to cooperation in such schemes.

It is generally thought that any form of associative interactions, such as those that are due to kinship²⁰, discrimination²¹ or "population viscosity"²², which includes spatial structure, would favour the evolution of cooperation (see refs 23, 24, for some exceptions). Such associations can lead to the formation of clusters of cooperators that can maintain cooperation against defecting invaders at the cluster boundaries²⁵. However, this mechanism does not operate in the spatial hawk–dove and snowdrift games. Ironically, the ultimate reason for this is that cooperation is already maintained in the well-mixed versions of these games, because the payoffs are such that it is best to adopt strategies that differ from the strategies of the opponents. This hinders cluster formation of cooperators in spatial populations.

Even though determination of payoff matrices in real systems is notoriously difficult, our results may be relevant for many natural populations. For example, predator inspection in sticklebacks is an often cited application of the Prisoner's Dilemma¹², but only the payoff ranking $T > R > S$ has been experimentally confirmed¹³. If P turns out to be less than S , predator inspection would actually be a snowdrift game. Similarly, RNA phages engage in Prisoner's Dilemma interactions in cells¹⁴, but selection alters the payoff structure, leading to stable coexistence of cooperating and defecting types in a snowdrift game²⁶. Other well-known examples of potential snowdrift or hawk–dove games include alarm calls in meerkat¹⁰ and fighting in large ungulates²⁷. Cooperation seems to be ubiquitous in meerkat, whereas serious escalations of fights seem to be common in musk ox. Because costs of alarm calls seem to be small¹⁰, whereas costs of forgoing reproduction are high, both observations are in agreement with our results that space should benefit cooperation for low cost-to-benefit ratios, but should lead to more frequent escalations for high ratios.

Overall, our results indicate that spatial extension of natural

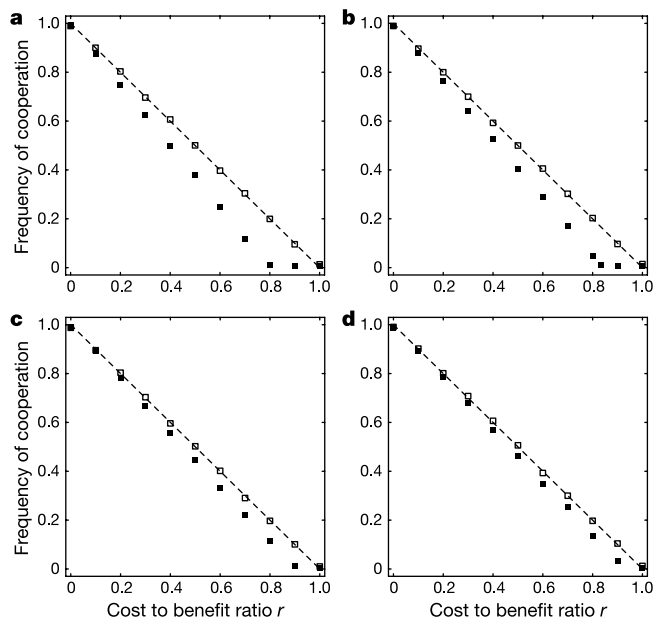


Figure 3 Average mixed strategy at stochastic equilibrium in the spatial hawk–dove game as a function of the parameter $r = \beta/\gamma$ for different lattice geometries. **a**, Triangular lattice, neighbourhood size $N = 3$; **b**, square lattice, $N = 4$; **c**, hexagonal lattice, $N = 6$; **d**, square lattice, $N = 8$. For synchronous updates (filled squares), spatial structure systematically increases the frequency of hawk-like behaviour as compared with well-mixed populations (dotted line). This effect becomes more pronounced for smaller N . By contrast, spatial structure barely affects the equilibrium strategy for asynchronous updates (open squares). See Methods for simulation details. Note that the equilibrium proportion of hawk-like behaviour in the mixed-strategy case is generally different from the equilibrium proportion of hawks in the pure strategy case (see Fig. 1). This contrasts with well-mixed populations, where the evolutionarily stable equilibrium is the same in both cases.

populations may decrease cooperation and lead to more frequent escalations of conflicts in situations in which cooperation persists in well-mixed populations. Thus, spatial structure may not be as universally beneficial for cooperation as previously thought. □

Methods

Spatial structure

In our spatially structured populations, individuals are confined to sites on regular 100×100 lattices with periodic boundary conditions, and interact with their neighbours. We used square lattices with $N = 4$ and $N = 8$ neighbours, hexagonal lattices ($N = 6$) and triangular lattices ($N = 3$). Whenever a site x is updated, a neighbour y is drawn at random among all N neighbours; the chosen neighbour takes over site x with probability $w_y = f(P_y - P_x)$, where the function f translates payoff differences into reproductive success, reflecting natural selection based on relative fitness. The site x remains unchanged, with probability $1 - w_y$. Lattice updating can be either synchronous or asynchronous. For synchronous updates, first all individuals interact in their respective neighbourhood and then all sites are updated simultaneously through competition with a randomly chosen neighbour. For asynchronous updates, only a single, randomly selected focal site is updated at each simulation step: first the payoffs of the focal individual and a random neighbour are determined, after which these two individuals compete to re-populate the focal site. See Supplementary Information for the case where competition involves all neighbours, rather than just a randomly chosen one.

Pure strategies

With pure strategies, each individual is either a cooperator or a defector. Lattices are initialized randomly with equal proportions of the two strategies. $f(z) = z_+/ \alpha$ determines the transition probabilities, where z_+ is equal to z if $z > 0$ and 0 otherwise, and where $\alpha = T - P$ in the snowdrift game and $\alpha = T - S$ in the Prisoner's Dilemma, ensuring that $f(P_y - P_x) \leq 1$. In well-mixed populations this implements the replicator dynamics¹⁷. Equilibrium frequencies of cooperators and defectors are obtained by averaging over 1,000 generations after a relaxation time of 5,000 generations.

The individual-based spatial models are complemented by deterministic pair-approximation (ref. 28 and see Supplementary Information). This approach correctly predicts a decrease in the frequency of cooperators in spatially structured populations, but it underestimates the effects of local correlations: for larger r the fragility of cooperative clusters is underrated, as is the ability of cooperators to displace defectors for small r (Fig. 1). Near the extinction thresholds, interesting symmetrical dynamics occur: tiny patches of defectors (cooperators) meander in a sea of cooperators (defectors). Occasionally they divide into pairs or collide and vanish. This resembles a branching and annihilating random walk, which suggests that there are critical phase transitions and points to interesting relationships between game theory and condensed matter physics²⁹.

Mixed strategies

For mixed strategies in the hawk–dove game, an individual is characterized by the probability p to show dove-like behaviour. Exploration of this continuous strategy space requires mutations. Whenever an individual with strategy p reproduces, a mutation occurs with a small probability (0.01) that assigns the offspring the strategy $p + \xi$, where ξ denotes a gaussian-distributed random variable with a mean of 0 and an s.d. of 0.002. To speed up simulations, the lattice is initialized with random strategies drawn from a normal distribution with a mean corresponding to the equilibrium strategy in well-mixed populations and an s.d. of 0.02. The simulation results are insensitive to the initialization details.

An individual in x with strategy p interacting with a neighbour with strategy q gets an average payoff $P_x = pqR + p(1 - q)S + (1 - p)qT + (1 - p)(1 - q)P$. The small difference in the strategies of parents and mutant offspring leads to small payoff differences $P_y - P_x$ between neighbouring individuals. Thus, the update rule for pure strategies returns small probabilities for a strategy change, which slows down the simulations. We therefore used the nonlinear function $f(z) = [1 + \exp(-z/\kappa)]^{-1}$, in which κ is a noise term that reflects uncertainties in assessing the payoffs. This nonlinearity greatly speeds up the simulations and introduces an interesting and realistic form of error, whereby a worse performing player occasionally manages to reproduce. For $\kappa \rightarrow \infty$, errors in assessing the payoffs increase until no information is left and the players randomly adopt neighbouring strategies. We used $\kappa = 0.1$ in our simulations. The equilibrium levels of dove-like behaviour were determined by evolving the lattice over 10,000 generations and then averaging the mixed strategies over another 1,000 generations.

Received 11 September 2003; accepted 21 January 2004; doi:10.1038/nature02360.

1. Maynard Smith, J. & Szathmáry, E. *The Major Transitions in Evolution* (W. H. Freeman, Oxford, UK, 1995).
2. von Neumann, J. & Morgenstern, O. *Theory of Games and Economic Behaviour* (Princeton Univ. Press, Princeton, 1944).
3. Maynard Smith, J. & Price, G. The logic of animal conflict. *Nature* **246**, 15–18 (1973).
4. Axelrod, R. & Hamilton, W. D. The evolution of cooperation. *Science* **211**, 1390–1396 (1981).
5. Sugden, R. *The Economics of Rights, Co-operation and Welfare* (Blackwell, Oxford, UK, 1986).
6. Nowak, M. A. & May, R. M. Evolutionary games and spatial chaos. *Nature* **359**, 826–829 (1992).
7. Doebeli, M. & Knowlton, N. The evolution of interspecific mutualisms. *Proc. Natl Acad. Sci. USA* **95**, 8676–8680 (1998).
8. Killingback, T., Doebeli, M. & Knowlton, N. Variable investment, the continuous prisoner's dilemma, and the origin of cooperation. *Proc. R. Soc. Lond. B* **266**, 1723–1728 (1999).
9. Wilkinson, G. S. Reciprocal food-sharing in the vampire bat. *Nature* **308**, 181–184 (1984).

10. Clutton-Brock, T. H. *et al.* Selfish sentinels in cooperative mammals. *Science* **284**, 1640–1644 (1999).
11. Seyfarth, R. M. & Cheney, D. L. Grooming alliances and reciprocal altruism in vervet monkeys. *Nature* **308**, 541–543 (1984).
12. Milinski, M. Tit for tat in sticklebacks and the evolution of cooperation. *Nature* **325**, 433–435 (1987).
13. Milinski, M., Lüthi, J. H., Eggler, R. & Parker, G. A. Cooperation under predation risk: experiments on costs and benefits. *Proc. R. Soc. Lond. B* **264**, 831–837 (1997).
14. Turner, P. E. & Chao, L. Prisoner's dilemma in an RNA virus. *Nature* **398**, 441–443 (1999).
15. Heinsohn, R. & Parker, C. Complex cooperative strategies in group-territorial African lions. *Science* **269**, 1260–1262 (1995).
16. Clutton-Brock, T. Breeding together: kin selection and mutualism in cooperative vertebrates. *Science* **296**, 69–72 (2002).
17. Hofbauer, J. & Sigmund, K. *Evolutionary Games and Population Dynamics* (Cambridge Univ. Press, Cambridge, UK, 1998).
18. Killingback, T. & Doebeli, M. Spatial evolutionary game theory: hawks and doves revisited. *Proc. R. Soc. Lond. B* **263**, 1135–1144 (1996).
19. Maynard Smith, J. *Evolution and the Theory of Games* (Cambridge Univ. Press, Cambridge, UK, 1982).
20. Hamilton, W. D. The evolution of altruistic behaviour. *Am. Nat.* **97**, 354–356 (1963).
21. Nowak, M. A. & Sigmund, K. Evolution of indirect reciprocity by image scoring. *Nature* **393**, 573–577 (1998).
22. Mitteldorf, J. & Wilson, D. S. Population viscosity and the evolution of altruism. *J. Theor. Biol.* **204**, 481–496 (2000).
23. Wilson, D. S., Pollock, G. B. & Dugatkin, L. A. Can altruism evolve in purely viscous populations? *Evol. Ecol.* **6**, 31–34 (1992).
24. Taylor, P. D. Altruism in viscous populations—an inclusive fitness approach. *Evol. Ecol.* **6**, 352–356 (1992).
25. Hauert, C. Fundamental clusters in spatial 2×2 games. *Proc. R. Soc. Lond. B* **268**, 761–769 (2001).
26. Turner, P. E. & Chao, L. Escape from prisoner's dilemma in RNA phage $\phi 6$. *Am. Nat.* **161**, 497–505 (2003).
27. Wilkinson, G. S. & Shank, C. C. Rutting-fight mortality among musk oxen on banks island, Northwest Territories, Canada. *Anim. Behav.* **24**, 756–758 (1977).
28. Matsuda, H., Ogita, N., Sasaki, A. & Sato, K. Statistical mechanics of populations. *Prog. Theor. Phys.* **88**, 1035–1049 (1992).
29. Szabó, G. & Hauert, C. Phase transitions and volunteering in spatial public goods games. *Phys. Rev. Lett.* **89**, 118101 (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank M. Ackermann and G. Szabó for comments. C.H. acknowledges support from the Swiss National Science Foundation. M.D. was supported by the National Science and Engineering Research Council (NSERC), Canada, and by the James S. McDonnell Foundation, USA.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to C.H. (hauert@zoology.ubc.ca).

Emergence of cooperation and evolutionary stability in finite populations

Martin A. Nowak^{1,2}, Akira Sasaki⁴, Christine Taylor^{1,5} & Drew Fudenberg³

¹Program for Evolutionary Dynamics, ²Department of Organismic and Evolutionary Biology and Department of Mathematics, ³Department of Economics, Harvard University, Cambridge, Massachusetts 02138, USA
⁴Department of Biology, Kyushu University, Fukuoka 812-8581, Japan
⁵Department of Mathematics, MIT, Cambridge, Massachusetts, 02139, USA

To explain the evolution of cooperation by natural selection has been a major goal of biologists since Darwin. Cooperators help others at a cost to themselves, while defectors receive the benefits of altruism without providing any help in return. The standard game dynamical formulation is the 'Prisoner's Dilemma'^{1–11}, in which two players have a choice between cooperation and defection. In the repeated game, cooperators using direct reciprocity cannot be exploited by defectors, but it is unclear how such cooperators can arise in the first place^{12–15}. In general, defectors are stable against invasion by cooperators. This understanding is based on traditional concepts of evolutionary stability and

dynamics in infinite populations^{16–20}. Here we study evolutionary game dynamics in finite populations^{21–25}. We show that a single cooperator using a strategy like ‘tit-for-tat’ can invade a population of defectors with a probability that corresponds to a net selective advantage. We specify the conditions required for natural selection to favour the emergence of cooperation and define evolutionary stability in finite populations.

In the Prisoner’s Dilemma, two players are offered a certain

Box 1 Game dynamics in finite populations

The fitness of strategies A and B with payoff matrix (1) is, respectively, given by:

$$\begin{aligned} f_i &= 1 - w + w[a(i-1) + b(N-i)]/[N-1] \\ g_i &= 1 - w + w[c i + d(N-i-1)]/[N-1] \end{aligned} \quad (4)$$

Here i denotes the number of individuals using strategy A, and $w \in [0, 1]$ specifies the contribution of the game to fitness. Selection dynamics can be formulated as a Moran process with frequency-dependent fitness. At each time step, an individual is chosen for reproduction proportional to its fitness. One identical offspring is being produced that replaces another randomly chosen individual. Thus N is strictly constant. The probability of adding an A-offspring is $if_i/[if_i + (N-i)g_i]$. At each time step, the number of A individuals can either increase by one, stay the same, or fall by one. Therefore, the transition matrix of the Markov process is tri-diagonal and defines a birth–death process given by:

$$\begin{aligned} P_{i,j+1} &= \frac{if_i}{if_i + (N-i)g_i} \frac{N-i}{N} \\ P_{i,j-1} &= \frac{(N-i)g_i}{if_i + (N-i)g_i} \frac{i}{N} \end{aligned} \quad (5)$$

We have $P_{i,i} = 1 - P_{i,j+1} - P_{i,j-1}$. All other entries of the transition matrix are 0.

The process has two absorbing states, $i = 0$ and $i = N$: if the population has reached either one of those states, then it will stay there forever. We denote by x_i the probability of ending up in state $i = N$ when starting in state i . We have:

$$x_i = P_{i,j+1}x_{i+1} + P_{i,i}x_i + P_{i,j-1}x_{i-1} \quad (6)$$

with boundary conditions $x_0 = 0$ and $x_N = 1$.

Let us calculate the probability, $\rho_A (= x_1)$, that a single individual A can invade and take over a population of B players. Solving equation (6), we obtain²⁷:

$$\rho_A = 1 / \left(1 + \sum_{k=1}^{N-1} \prod_{i=1}^k \frac{g_i}{f_i} \right) \quad (7)$$

If $\rho_A > 1/N$, then selection favours A replacing B.

The rate of evolution from all-B to all-A is given by $r = N\rho_A u$, where u is the mutation rate. We can rescale the rate of evolution in units of u . Thus, we set $u = 1$. The rate of evolution, r , can be an increasing or decreasing function of w . There can also be a maximum or minimum value of r for some intermediate $w \in (0, 1)$. It is possible that $r > 1$ for small w , but $r < 1$ for large w , or vice versa. For $w = 0$, we have $r = 1$.

In the limit of weak selection, $w \ll 1$, we find that:

$$N\rho_A \approx 1/[1 - (\alpha N - \beta)(w/6)] \quad (8)$$

with $\alpha = a + 2b - c - 2d$ and $\beta = 2a + b + c - 4d$. From this equation, we see that $N\rho_A > 1$ if $\alpha N > \beta$, which leads to equation (2). If $\alpha > 0$ then there is a minimum N for which $N\rho_A > 1$. It is given by $N_{\min} = \beta/\alpha$.

We define a strategy to be an ESS_N if selection opposes the invasion and fixation of any other strategy for any $w > 0$. Thus a necessary condition for ESS_N is that $N\rho_A$, as given by equation (8), is less than one. If instead we want to know whether a strategy is evolutionarily stable for a given N and a given w , then we cannot necessarily use the simplified equation (8), which only holds in the limit of small w . Instead, we have to check that the exact expression $N\rho_A$, as given by equation (7), is less than one.

payoff, R , for mutual cooperation and a lower payoff, P , for mutual defection. If one player cooperates while the other defects, then the cooperator gets the lowest payoff, S , while the defector gains the highest payoff, T . Thus, we have $T > R > P > S$. In the non-repeated Prisoner’s Dilemma, defectors dominate cooperators, which means that in any mixed population, defectors have a higher fitness. In the repeated Prisoner’s Dilemma, the same two players meet more than once, and there are many conceivable strategies that allow cooperative behaviour which cannot be defeated by defectors. The most famous such strategy is tit-for-tat (TFT), in which the player cooperates in the first round and then does whatever the opponent did in the previous round. Another strategy is always to defect (AllD). If the number of rounds is sufficiently large, then AllD and TFT resist invasion attempts by the other strategy. Thus, TFT can maintain cooperation, but how it can become established is unclear.

In the standard evolutionary model of the finitely repeated Prisoner’s Dilemma, TFT cannot invade AllD, because if everyone in an infinitely large population uses AllD, then a small fraction of TFT players have a lower payoff. Every invasion attempt by TFT is therefore eliminated by natural selection. Past work has proposed several modifications to this model that allow TFT to invade successfully: (1) a mass of TFT players arises simultaneously to overcome an invasion barrier¹²; (2) TFT players form spatial clusters^{13,14}; or (3) aggregate payoffs are stochastic¹⁵. Here we show that none of these modifications are necessary to explain the emergence of cooperation in finite populations.

Consider a game between two strategies, A and B, with payoff matrix:

$$\begin{array}{cc} & \begin{array}{c} A \quad B \end{array} \\ \begin{array}{c} A \\ B \end{array} & \begin{pmatrix} a & b \\ c & d \end{pmatrix} \end{array} \quad (1)$$

If A and B denote, respectively, TFT and AllD, then we have $a > c > d > b$ provided the number of rounds is finite and greater than $(T - P)/(R - P)$. In this case, both TFT and AllD are strict Nash equilibria and evolutionarily stable strategies (ESS).

Box 2 Invading AllD

Let strategies A and B denote, respectively, TFT and AllD in a Prisoner’s Dilemma which is repeated for n rounds on average. The payoff matrix is $a = Rn$, $b = S + P(n-1)$, $c = T + P(n-1)$ and $d = Pn$. If the average number of rounds, n , exceeds $(T - P)/(R - P)$ then we have $a > c > d > b$. In this case, there is an unstable equilibrium between TFT and AllD, and neither strategy can invade the other in the context of deterministic dynamics of infinitely large populations.

Condition (2) implies $n(R - P)(N - 2) > T(N + 1) - S(2N - 1) + P(N - 2)$. This inequality determines the minimum number of rounds required for (weak) selection to favour TFT replacing AllD for a given N . We need at least $N = 3$. For large N , the number of rounds must fulfill $n > (T + P - 2S)/(R - P)$. Let $R = 3$, $T = 5$, $P = 1$, $S = 0$. For $N = 3$ we have $n > 10.5$. For $N = 4$ we have $n > 6.75$. For large N we have $n > 3$.

Among all strategies of the repeated Prisoner’s Dilemma, TFT maximizes the probability of invading AllD in any finite population. More precisely, TFT belongs to a set of strategies that maximize this probability. All strategies of this set have the following property: (1) when playing themselves they cooperate all the time; (2) when playing AllD they cooperate on the first move and then never again. ‘Win-stay, lose-shift’ and ‘generous tit-for-tat’ have lower invasion probabilities because they attempt to cooperate with AllD too often. They work best once cooperation has been established^{12,30}.

Deterministic replicator dynamics of infinite populations admit an unstable equilibrium at a frequency of TFT given by $x^* = (d - b)/(a - b - c + d)$. If the initial frequency of TFT is less than this value, then it will be eliminated by natural selection. TFT can only replace AllD if its initial frequency exceeds this invasion barrier. The same evolutionary dynamics hold for AllD competing with other cooperative strategies such as 'generous tit-for-tat', 'contrite tit-for-tat' or 'win-stay, lose-shift'.

We now study evolutionary game dynamics in finite populations. In this case, there is always a non-zero probability that a mutant strategy can invade and take over the population even though it is opposed by selection. TFT invading AllD has a lower fitness if its frequency is less than x^* but has a higher fitness if its frequency is greater than this threshold. What is the probability that a single TFT player will take over a population of AllD players?

Consider a stochastic process describing a finite population of size N . At each time step, one individual is chosen for reproduction proportional to its fitness, and its offspring replaces a randomly chosen individual²⁶. The population size is constant. The fitness of each player depends on the number of TFT or AllD players. In addition, we introduce a parameter, w , which determines the contribution of the game's payoff to fitness. This parameter, quantifying the intensity of selection, cancels itself out in deterministic replicator dynamics of infinite populations, but plays a crucial role in the stochastic process that we study here.

We calculate the probability, ρ_A , that a single individual using strategy A will invade and take over a population of B players²⁷ (Box 1). For a neutral mutant²⁸ this fixation probability is $\rho_A = 1/N$. If $\rho_A > 1/N$ then selection favours A replacing B. In Fig. 1, we show that in the case of TFT and AllD, $N\rho_A$ is a one-humped function of N . For a wide choice of parameter values a, b, c, d and w there is an intermediate range of population sizes which ensure $N\rho_A > 1$. Thus, the invasion and replacement of AllD by TFT, starting from a single individual of TFT, can be favoured by natural selection. Interestingly, there are critical minimum and maximum population sizes that allow positive selection of TFT. In very small populations, there is a strong effect of spite: helping another individual leads to a significant disadvantage. For example, in a population of only two players, TFT always has a lower fitness than AllD. In a very large population it is extremely unlikely to reach the invasion barrier when starting with a single TFT player. Thus, neither small nor large but intermediate population sizes are optimum for initiating cooperation.

Can we derive the underlying principle that determines whether a particular payoff matrix (1) allows selection for TFT replacing

AllD? The exact expression for ρ_A is complicated. The condition $\rho_A > 1/N$ requires the solution of N th-order polynomials, and a diffusion approximation yields transcendental equations. Nevertheless, the following surprisingly simple theorem holds. For a given N and sufficiently weak selection (small w), selection favours TFT replacing AllD if:

$$a(N-2) + b(2N-1) > c(N+1) + d(2N-4) \quad (2)$$

For the smallest possible population size, $N = 2$ (it takes two to play), inequality (2) yields $b > c$ (which is not possible for the game between TFT and AllD). For the limit of large N , we obtain $a + 2b > c + 2d$. The latter condition is equivalent to $x^* < 1/3$. Therefore, if the invasion barrier of TFT is less than $1/3$, there can be positive selection for TFT to replace AllD in a finite population (Box 2).

In general, for any two strategies which are the best replies to themselves, we find that selection can favour A replacing B for some N and w , if $b > c$ or $x^* < 1/3$ (Fig. 2).

Our results have immediate consequences for the concept of evolutionary stability. The well-known definition of an ESS is motivated by selection dynamics in infinite populations^{16–20}. Strategy B is an ESS if either (1) $d > b$ or (2) $d = b$ and $a < c$. These conditions imply that selection opposes the spread of infinitesimally small fractions of A in infinitely large populations of B.

For finite N , we propose that B is an ESS, denoted ESS_N , if two conditions hold: (1) selection opposes A invading B, which means that a single mutant A in a population of B has a lower fitness; and (2) selection opposes A replacing B, which means $\rho_A < 1/N$, for any $w > 0$. The first condition is equivalent to $b(N-1) < c + d(N-2)$. The second condition, for small w , is equivalent to $a(N-2) + b(2N-1) < c(N+1) + d(2N-4)$. For $N = 2$, both conditions reduce to $b < c$. For large populations, the two conditions lead to $b < d$ and $x^* > 1/3$, respectively. Hence, for small populations, the traditional ESS concept is neither necessary nor sufficient; for large populations, it is necessary but not sufficient (Fig. 3). If we consider a game with many different strategies, then the two conditions must hold in pairwise comparison with every other strategy.

The motivation of the ESS_N concept is as follows. If a strategy is an ESS_N , then a single mutant of any other strategy must have a lower fitness. Therefore, selection opposes the initial spread of any other strategy. As we have seen in the case of AllD and TFT, however, in a finite population it is possible that the fixation of a strategy is favoured by selection even though its initial increase is opposed by selection. Thus, the second condition states that a strategy can only be an ESS_N if the fixation probability of all other strategies is less

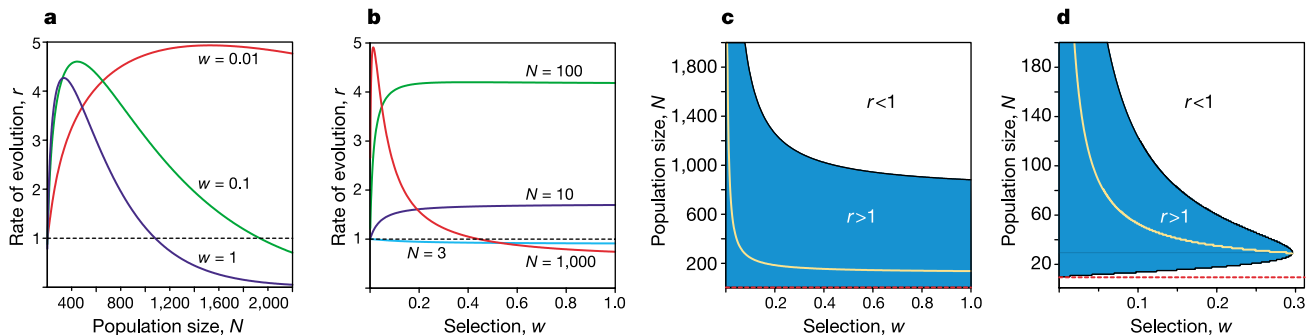


Figure 1 Selection can favour the replacement of AllD by TFT in finite populations. **a**, The rate of evolution, $r = N\rho_A$, is a one-humped function of N . There is an intermediate range of N which leads to positive selection of TFT, $N\rho_A > 1$. **b**, $N\rho_A$ is shown as function of w . For small N , we have $N\rho_A < 1$ for all w . For larger N we have $N\rho_A > 1$ for all w . For even larger N we have $N\rho_A > 1$ as long as w is below a certain threshold. **c, d**, The blue-

shaded region indicates the parameter region where $N\rho_A > 1$. The yellow lines show the optimum value N for a given w maximizing $N\rho_A$. The broken red line indicates $N_{\min} = (2a + b + c - 4d)/(a + 2b - c - 2d)$, which is the predicted minimum population size required for positive selection of TFT in the limit of weak selection. Parameter choices are: $R = 3$, $T = 5$, $P = 1$, $S = 0$; and $n = 10$ rounds for **a–c** and $n = 4$ rounds for **d**.

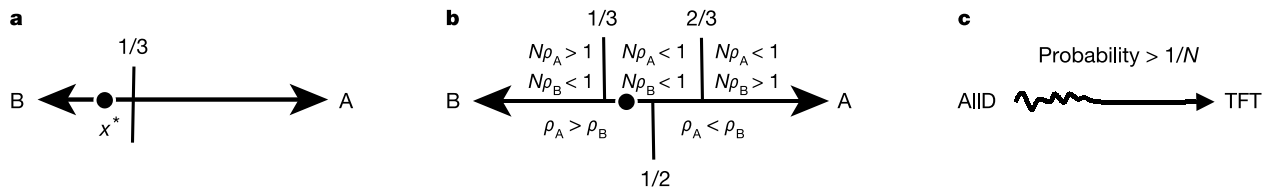


Figure 2 The 1/3-law of frequency-dependent evolution. **a**, Suppose A and B are the best replies to themselves, meaning $a > c$ and $d > b$ in payoff matrix (1). In this case all-A and all-B are stable equilibria of the replicator dynamics for infinite population size. The unstable equilibrium is located at a frequency of A given by $x^* = (d - b)/(a - b - c + d)$. If $x^* < 1/3$ then selection favours A replacing B for a sufficiently large population and weak selection. The minimum population size is given by $N_{\min} = (2a + b + c - 4d)/(a + 2b - c - 2d)$. **b**, The relationship between the 1/3-law and risk dominance. If $x^* < 1/3$ then $\rho_A > 1/N > \rho_B$. If $1/3 < x^* < 1/2$ then $1/N > \rho_A > \rho_B$. If

$1/2 < x^* < 2/3$ then $1/N > \rho_B > \rho_A$. If $2/3 < x^*$ then $\rho_B > 1/N > \rho_A$. We note that if A is risk dominant, meaning $x^* < 1/2$, then $\rho_A > \rho_B$. These results hold in the limit of weak selection and large population size. The location of x^* as shown implies that $1/N > \rho_A > \rho_B$. **c**, AIID and TFT in the repeated Prisoner's Dilemma. Although AIID is evolutionarily stable against invasion by TFT for deterministic dynamics of infinite populations, in a finite population the probability that a single mutant of TFT takes over an AIID population can exceed $1/N$. Selection can favour TFT replacing AIID, if the unstable equilibrium occurs at a frequency of less than $1/3$ TFT.

than the neutral threshold, $1/N$. In summary, we simply ask that a homogeneous ESS_N population is protected by selection against invasion and replacement. These requirements represent a natural extension of the original ESS concept formulated by Maynard Smith for infinitely large populations and deterministic evolutionary dynamics¹⁸.

Schaffer²² has proposed that a strategy is evolutionarily stable in a finite population if a single mutant of any other strategy has lower fitness. This is the first of our two conditions. Schaffer²² also proposes a global stability condition, namely that the ESS strategy must have a greater fitness than the other strategy for any composition of the population. This very stringent condition is a finite- N extension of Hamilton's unbeatable strategy²⁹, which dominates every other strategy. We note that every unbeatable strategy is an ESS, but the reverse is not true. Unfortunately, unbeatable strategies are rare. Many biological games admit ESS, but not unbeatable strategies.

Sometimes it is of interest to ask whether strategy A is more likely to replace B than vice versa. Let ρ_A and ρ_B denote the respective fixation probabilities. In the case where both A and B are the best replies to themselves and in the limit of weak selection and large populations, we find that $\rho_A > \rho_B$ if A is risk-dominant, meaning that $a + b > c + d$. For general N and w , however, risk dominance does not determine the ordering of the fixation probabilities.

a		A	B	A is an ESS_N for $N > 12$ B is an ESS_N for $N < 53$ $\rho_A > \rho_B$ for $N > 19$
	A	20	0	
	B	17	1	
b		A	B	A is an ESS_N for $N < 22$ B is an ESS_N for $N > 17$ $\rho_A > \rho_B$ for $N < 20$
	A	1	28	
	B	2	30	

Figure 3 A strategy is ESS_N if it is protected by selection against invasion and replacement by another strategy for given N and any $w > 0$. **a**, Both A and B are classical ESS, but for $2 \leq N \leq 12$ only B is ESS_N ; for $12 < N < 53$ both A and B are ESS_N , for $N \geq 53$ only A is ESS_N . We note that strategy B is a classical ESS, but is not ESS_N for large N ; in large populations there is selection for A replacing B. **b**, B dominates A. Therefore only B is a classical ESS. For $2 \leq N \leq 17$, however, we obtain that only A is ESS_N . For $17 < N < 22$ both A and B are ESS_N . For $N \geq 22$ only B is ESS_N . Examples **a** and **b** illustrate that for small populations the traditional ESS concept is neither necessary nor sufficient to imply ESS_N , and for large populations it is necessary but not sufficient.

Here we have studied a frequency-dependent Moran process, which is a natural finite- N analogue to the replicator equation. One can envisage many different stochastic processes that describe game dynamics in finite populations. An interesting possibility is the following. Pick two players at random. One is chosen for reproduction, the other for elimination. Hence, only mixed pairs can change the population. Suppose that player A is chosen for reproduction with probability $f_i/(f_i + g_i)$ and player B with probability $g_i/(f_i + g_i)$. In this case, we obtain the same process as we have analysed here, up to rescaling time.

If, instead, the fitter player is always chosen for reproduction, then the resulting process is stochastic in speed, but deterministic in direction: it will always follow the gradient of selection. But if player A is chosen for reproduction with probability $1/(1 + \exp[-(f_i - g_i)/\tau])$, then parameter w cancels out. There is, however, a new parameter, τ , which has a similar role and is equivalent to temperature in statistical physics. If $\tau \rightarrow 0$ then the fitter player is always chosen; selection is strong. If $\tau \rightarrow \infty$ then selection is weak, and the process is dominated by random drift. In the limit of large τ , we obtain exactly the same results as are presented here.

Another possibility is studying a frequency-dependent Wright-Fischer process with discrete generations. Furthermore, in all of those models, stochasticity could arise in evaluating the payoffs of individual players. We expect that all these processes (as long as they are not deterministic in following selection) will have a similar behaviour to what we have described here.

To sum up, (1) in finite populations, natural selection can favour the invasion and replacement of the AIID strategy by a cooperative strategy, when starting from a single individual using that strategy. No specific mechanism of invasion is required. (2) For any two strategies A and B, natural selection can favour A replacing B in a finite population provided $b > c$ or $a - c > 2(d - b)$. If A and B are the best replies to themselves then the latter condition implies that the frequency of A at the unstable equilibrium, x^* , must be less than $1/3$. (3) Our analysis leads to natural conditions for evolutionary stability in finite populations. These conditions specify whether a given resident strategy is protected by selection against invasion and replacement of any mutant strategy. $\square$

Methods

Remarks on ESS

If $d > b$ then B is both a strict Nash equilibrium and an ESS in comparison with A. A strict Nash equilibrium implies protection by selection against replacement in the following sense: for a given payoff matrix (a, b, c, d) with $d > b$ and for any given intensity of selection, $0 < w \leq 1$, we have $\rho_A \rightarrow 0$ as $N \rightarrow \infty$.

For every finite population size, N , however, we can calculate the maximum net selective advantage for a mutant replacing a strict Nash equilibrium. Given b, d with $d > b$, what is the maximum probability ρ_A of A replacing B? We are free to choose a and c .

To maximize ρ_A , we set $a \rightarrow \infty$ and $c = 0$. For large populations, we obtain $\rho_A = [1 - w(1 - b)]/[2 - w(2 - b - d)]$. For $w \rightarrow 0$ we have $\rho_A = 1/2$. For $w = 1$ we have $\rho_A = b/(b + d)$. This fixation probability of A corresponds to a constant relative fitness of $1 + (b/d)$ or a net selective advantage of b/d . Hence there can be enormous selection pressure for replacement of a strict Nash equilibrium in arbitrarily large, finite populations (when the other equilibrium is much more efficient).

The calculation here uses the fact that from state $i = 1$ the system can go either to $i = 0$ or $i = 2$. Because $a \rightarrow \infty$ and $c = 0$, fixation of strategy A is certain from state $i = 2$. Hence, the fixation probability from $i = 1$ to $i = N$ is just the probability $P_{12}/(P_{12} + P_{10}) = (1 - w + wb)/(1 - w + wb + 1 - w + wd(N - 2)/(N - 1))$. This holds for all w . For large N , we obtain the above formula for ρ_A .

Risk dominance

Let ρ_A denote the probability that a single A player reaches fixation in a population of B. Let ρ_B denote the probability that a single B player reaches fixation in a population of A. We obtain:

$$\frac{\rho_A}{\rho_B} = \prod_{i=1}^{N-1} \frac{f_i}{g_i} \quad (3)$$

For weak selection (small w) we find $\rho_A/\rho_B = 1 + w[(N/2)(a + b - c - d) + d - a]$. It follows that $\rho_A > \rho_B$ is equivalent to $(N - 2)(a - d) > N(c - b)$. For large N this means $a - c > d - b$. Hence, if both A and B strategies are strict Nash equilibria then the risk-dominant equilibrium has a higher fixation probability when starting from a single player using that strategy. For general N and w , risk-dominance does not decide whether ρ_A is greater than ρ_B .

More general strategies

We have mostly studied the dynamics between ALLD and TFT. The repeated Prisoner's Dilemma, like other repeated games, admits a huge set of possible strategies, which makes it difficult to explicitly analyse the dynamics of evolution. In general, a strategy for playing the repeated Prisoner's Dilemma is a mapping from any history of the game between two players into the interval $[0, 1]$, denoting the probability of cooperation on the next move. However, we note that for the finitely repeated game, ALLD is a strict Nash equilibrium in comparison with all cooperative strategies, where we define a 'cooperative strategy' as a strategy which cooperates on the first move. Let us divide cooperative strategies into two subsets: (1) those that are dominated by ALLD and (2) those that are bistable with ALLD. In an infinitely large population, no cooperative strategy can ever invade ALLD. In a finite population of size N , strategies of the second subset can invade and replace ALLD provided inequality (2) holds and selection is sufficiently weak.

In an infinitely repeated Prisoner's Dilemma with time-average payoffs, it turns out that TFT dominates ALLD. In this case it can be shown that the 'win-stay, lose-shift'³⁰ strategy (also known as 'Pavlov' or 'perfect tit-for-tat') is the only simple strategy which cannot be invaded by any other strategy, and that it is the only strategy that is evolutionarily stable in an infinite population when actions are taken with a vanishingly small probability of error³¹. Moreover, this strategy is also the unique ESS in a model where strategies are encoded by finite-state automata, and the complexity of the automaton represents an evolutionary cost³¹.

Received 10 December 2003; accepted 11 February 2004; doi:10.1038/nature02414.

- Trivers, R. The evolution of reciprocal altruism. *Q. Rev. Biol.* **46**, 35–57 (1971).
- Axelrod, R. & Hamilton, W. D. The evolution of cooperation. *Science* **211**, 1390–1396 (1981).
- Axelrod, R. *The Evolution of Cooperation* (Basic Books, New York, 1984).
- Milinski, M. Tit for tat in sticklebacks and the evolution of cooperation. *Nature* **325**, 433–435 (1987).
- May, R. M. More evolution of cooperation. *Nature* **327**, 15–17 (1987).
- Dugatkin, L. A. *Cooperation Among Animals* (Oxford Univ. Press, Oxford, UK, 1997).
- Pfeiffer, T., Schuster, S. & Bonhoeffer, S. Cooperation and competition in the evolution of ATP-producing pathways. *Science* **292**, 504–507 (2001).
- Fehr, E. & Fischbacher, U. The nature of human altruism. *Nature* **425**, 785–791 (2003).
- Hammerstein, P. (ed.) *Genetic and Cultural Evolution of Cooperation* (MIT Press, Cambridge, Massachusetts, 2003).
- Boyd, R., Gintis, H., Bowels, S. & Richerson, P. J. The evolution of altruistic punishment. *Proc. Natl Acad. Sci. USA* **100**, 3531–3535 (2003).
- Limburg, D. & Maskin, E. Evolution and cooperation in noisy repeated games. *Am. Econ. Rev.* **80**, 274–279 (1990).
- Nowak, M. A. & Sigmund, K. Tit for tat in heterogeneous populations. *Nature* **355**, 250–253 (1992).
- Nowak, M. A. & May, R. M. Evolutionary games and spatial chaos. *Nature* **359**, 826–829 (1992).
- Killingback, T. & Doebeli, M. Self-organized criticality in spatial evolutionary game theory. *J. Theor. Biol.* **191**, 335–340 (1998).
- Fudenberg, D. & Harris, C. Evolutionary dynamics with aggregate shocks. *J. Econ. Theor.* **57**, 420–441 (1992).
- Maynard Smith, J. & Price, G. R. Logic of animal conflict. *Nature* **246**, 15–18 (1973).
- Taylor, P. D. & Jonker, L. B. Evolutionary stable strategies and game dynamics. *Math. Biosci.* **40**, 145–156 (1978).
- Maynard Smith, J. *Evolution and the Theory of Games* (Cambridge Univ. Press, Cambridge, UK, 1982).
- Hofbauer, J. & Sigmund, K. *Evolutionary Games and Population Dynamics* (Cambridge Univ. Press, Cambridge, UK, 1998).
- Hofbauer, J. & Sigmund, K. Evolutionary game dynamics. *Bull. Am. Math. Soc.* **40**, 479–519 (2003).
- Riley, J. G. Evolutionary equilibrium strategies. *J. Theor. Biol.* **76**, 109–123 (1979).
- Schaffer, M. Evolutionary stable strategies for a finite population and a variable contest size. *J. Theor. Biol.* **132**, 469–478 (1988).
- Fogel, G., Andrews, P. & Fogel, D. On the instability of evolutionary stable strategies in small populations. *Ecol. Model.* **109**, 283–294 (1998).
- Ficci, S. & Pollack, J. *Effects of Finite Populations on Evolutionary Stable Strategies*. *Proc. 2000 Genetic and*

- Evolutionary Computation Conf.* (ed. Whitley, D.) 927–934 (Morgan-Kaufmann, San Francisco, 2000).
- Schreiber, S. Urn models, replicator processes, and random genetic drift. *Siam. J. Appl. Math.* **61**, 2148–2167 (2001).
- Moran, P. A. P. *The Statistical Processes of Evolutionary Theory* (Clarendon, Oxford, UK, 1962).
- Karlin, S. & Taylor, H. M. *A First Course in Stochastic Processes* 2nd edn (Academic, London, 1975).
- Kimura, M. Evolutionary rate at the molecular level. *Nature* **217**, 624–626 (1968).
- Hamilton, W. D. Extraordinary sex ratios. *Science* **156**, 477–488 (1967).
- Nowak, M. A. & Sigmund, K. A strategy of win-stay, lose-shift that outperforms tit for tat in Prisoner's Dilemma. *Nature* **364**, 56–58 (1993).
- Binmore, K. & Samuelson, L. Evolutionary stability in repeated games played by the finite automata. *J. Econ. Theor.* **57**, 278–305 (1992).

Acknowledgements The Program for Evolutionary Dynamics is supported by J. Epstein.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.A.N. (martin_nowak@harvard.edu).

Pre-social benefits of extended parental care

Jeremy Field & Selina Brace

Department of Biology, University College London, Wolfson House, 4 Stephenson Way, London NW1 2HE, UK

The evolution of helping, in which some individuals forfeit their own reproduction and help others to reproduce, is a central problem in evolutionary biology. Recently proposed insurance-based mechanisms rely on a pre-existing life history with a long period of offspring dependency relative to the short life expectancies of adult carers^{1–4}: a lone mother's offspring are doomed if she dies young, whereas after a helper dies, other group members can finish rearing the offspring^{5,6}. A critical question, however, is how this life history could evolve in ancestral non-social populations, as offspring survival would then depend on a single, short-lived carer. Here, we resolve this paradox by focusing on the extended parental care inherent in prolonged dependency. We show experimentally that in non-social wasps, extended care can significantly reduce the impact of interspecific parasites. Under extended care, offspring are less vulnerable by the time they are exposed to parasites, and costs of parasitism are reduced because mothers have the option to terminate investment in failing offspring. By experimentally simulating aspects of extended care in a species where it is lacking, we demonstrate that neither benefit requires specialized behaviour. Such benefits could therefore offset the disadvantage of prolonged dependency in non-social species, thereby facilitating the evolution of helping.

Immature nest-building wasps are helpless larvae that are entirely dependent on adult carers for food. The duration of parental care is minimized in 'mass provisioning' wasps, including most non-social taxa^{7,8}: before it even hatches from the egg, each offspring is sealed into a cell containing all of the food that it will require for maturation, so that it is nutritionally independent of its mother. In contrast, almost all eusocial and a few non-social wasps have extended parental care. These 'progressive provisioners' feed each developing larva gradually as it grows^{7,8}. Whereas a single mass provisioner can fully provision about 1 offspring per day, provisioning is extended over 5–70 days in progressive provisioners^{4,9,10}. Even if they provision multiple offspring simultaneously, non-social progressive provisioners will, on average, leave fewer independent offspring than mass provisioners, because mothers are more likely

to die before provisioning is complete (Fig. 1). If extended parental care evolved before helping, as assumed by many models^{1–4}, what benefits offset this obvious disadvantage?

The non-social wasp genus *Ammophila* (Sphecidae) has long been a focus of study^{11,12}, and includes species with and without extended care. Females nest alone and place each offspring in a separate burrow, which is kept closed when the female is away hunting. Mothers provide paralysed caterpillars as food (Fig. 2), laying an egg on the first caterpillar at the time it is placed in the burrow. Some species mass provision one offspring at a time, sealing each burrow permanently 1–2 days after oviposition^{12,13}. Other species, however, are progressive provisioners. These oviposit on the first caterpillar, but provide further food only gradually once the egg has hatched, so that provisioning is extended over 5–9 days^{9,12}. During periods between feeds, progressive provisioners start off other burrows, and usually have 2–4 offspring in mid-provisioning at a time⁹. Cuckoo flies (Diptera: Miltogramminae) are major natural enemies of wasps^{14–16}. They gain access to *Ammophila* burrows only while they are open during feeding events. Flies deposit live maggots that destroy the immature wasp and then eat the provisions^{9,13,17}.

We tested two mechanisms through which extended care might ameliorate the effects of interspecific parasitism. First, progressive provisioning might delay offspring being exposed to parasitism until they are less vulnerable. Each provisioning event represents an opportunity for parasites, but under progressive provisioning only the first event occurs before offspring enter the larval stage^{14,18}, whereas under mass provisioning an offspring receives all of its feeds while still an egg. In experiment 1 we tested whether larvae are better defended than eggs by challenging immature wasps of different ages with miltogrammine maggots (*Metopia*). In the progressive provisioner *Ammophila pubescens* Curtis⁹, 15 out of 23 wasp eggs were destroyed by maggots, whereas wasp larvae were almost immune to attack (only 3 out of 22 failed: $\chi^2 = 10.4$, $P < 0.005$). We found that offspring of the mass provisioner *Ammophila sabulosa* (L.)^{13,19} would be similarly protected were it to provision progressively: its larvae were also much more likely than eggs to survive experimental parasitism (25 out of 36 versus 2 out of 25 survived: $\chi^2 = 20.2$, $P < 0.001$). *A. sabulosa* eggs rarely hatch before provisioning is complete¹³, so that larvae are unlikely

to encounter fly maggots in nature—larval immunity probably results from unspecialized traits such as increased cuticular toughness.

A second possible advantage of progressive provisioning is that by investing only gradually as their offspring develop, mothers have the opportunity to intervene or terminate investment early if offspring show signs of failing²⁰. By the time a progressively provisioned burrow receives its second feed, cuckoo parasites introduced with the first feed will have killed the wasp egg and begun growing. In experiment 2, we tested whether mothers can respond adaptively to growing parasites⁹. We removed egg-bearing caterpillars from *A. pubescens* field nests on the evening of oviposition (Fig. 2b) and incubated them with fly maggots in the laboratory overnight. We replaced caterpillars in their burrows before wasp activity began next morning, by which time they bore growing maggots on the collapsed wasp egg. All 16 mothers that we tested abandoned their parasitized offspring after their next visit to the nest, thus saving approximately 80% of the provisioning effort that would have been wasted on a failed offspring. In contrast, 14 out of 19 controls had further provisioned their nests by the time our observations ended ($\chi^2 = 16.7$, $P < 0.001$).

Experiment 2 showed that *A. pubescens* mothers respond adaptively to parasites, but it did not test whether their response depends on extended parental care. This would be the case if mothers were able to respond only to parasites that have grown during the delays between feeds that occur under progressive provisioning. In experiment 3 we performed this critical test using the mass provisioner *A. sabulosa*. Soon after a mother had oviposited on her first caterpillar, we removed the nest contents and either added five tiny fly maggots, leaving the wasp egg intact (treatment 1), or added one to two larger maggots and squashed the wasp egg (treatment 2). Treatment 1 is what a mass provisioner would typically encounter at a parasitized

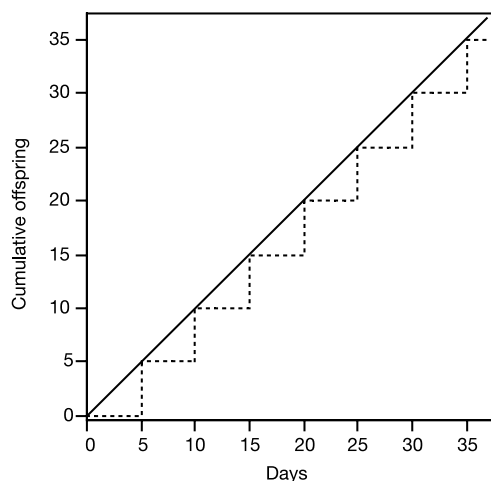


Figure 1 Offspring production under mass versus progressive provisioning. The graph shows the cumulative number of offspring fully provisioned by a non-social mass provisioner (solid line) that can fully provision one offspring per day, and by a progressive provisioner (dashed line) that uses the same food to provision batches of five offspring simultaneously over 5-day periods. On average, the mass provisioner will die with more of her past investment represented by fully provisioned offspring.

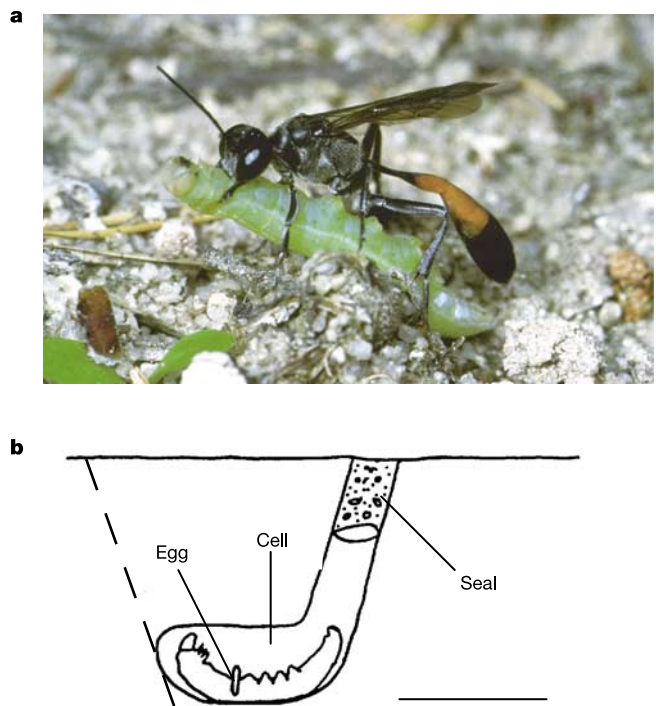


Figure 2 *Ammophila* nesting biology. **a**, *Ammophila pubescens* female carrying a paralysed caterpillar to her burrow. (Photograph by M. Blösch.) **b**, *Ammophila* burrow with first caterpillar. Soil to the left of the dashed line was excavated in experiments 2–3, allowing contents to be removed without damaging the burrow or its entrance. Scale bar, 2 cm.

nest, whereas treatment 2 mimics what *A. sabulosa* would encounter after a feeding delay at the evolutionary origin of progressive provisioning. Because *A. sabulosa* normally brings additional feeds within hours of oviposition, the maggots did not grow appreciably before the mothers encountered them on their next visit to the nest. All nine mothers exposed to tiny maggots showed no detectable response, and continued provisioning their doomed offspring to completion. In contrast, five out of twelve mothers exposed to larger maggots abandoned their burrows on the next visit, after pulling out the burrow contents (Fisher's exact $P = 0.045$). A further six of these mothers re-oviposited after pulling the manipulated caterpillar from the burrow and either discarding it, or attempting to remove fly maggots and then replace it in the burrow. Decisions to abandon appeared to be strategic: females whose burrows had contained a greater weight of food when maggots were added tended to be less likely to abandon (U -test, $P = 0.07$). The response to growing parasites may reflect periodic exposure to them during mass provisioning; at some sites up to 40% of *A. sabulosa* nests received their final feeds on the day after oviposition (J.F. and S.B., unpublished data), by which time parasites will have grown and destroyed the wasp egg.

Our results indicate how extended parental care might evolve in non-social populations, potentially facilitating the subsequent evolution of helping^{1–6}. Extended care provides two complementary advantages. The first is that offspring are older, and therefore less vulnerable^{21–23}, when they are exposed to enemies. The second advantage is that mothers are in a position to detect mortality factors that become apparent only gradually, and can avoid wasting a full quota of investment on affected offspring. Such mortality factors^{16,17} might include offspring predation, larval diseases with delayed symptoms, or fungi attacking food provisions, as well as cuckoo parasites. For a mass provisioner, however, it is too late, as females have already invested fully in their offspring by the time that such mortality becomes detectable.

Extended parental care magnifies insurance-based advantages that favour helping, and its distribution thus helps to define an important dichotomy in how sociality evolved²⁴. Progressive provisioning occurs sporadically among wasps and bees, including most major groups with eusocial species^{12,14,25–27}. Its absence from other groups, notably halictid bees²⁶, may partly reflect a constraint: the ancestral sequence of oviposition and provisioning. Both non-social and social halictids oviposit only at the end of provisioning²⁶, whereas progressive provisioning requires the reverse sequence. The occurrence of progressive provisioning will also depend on the balance between costs through prolonging offspring dependency versus benefits that depend on the frequency and type of mortality factors operating. High maternal mortality rates and larger numbers of offspring provisioned simultaneously will lead to increased costs (see Fig. 1). Greater prevalence of mortality factors whose effects are mitigated through progressive provisioning will lead to increased benefits. Our results show how, even at its evolutionary origin, progressive provisioning can lessen significantly the impact of cuckoo parasitism, a major source of immature mortality in key non-social taxa such as eumenine wasps, the sister group of social vespids^{25,28}. Similar advantages are likely to apply for other kinds of enemy^{16,17,21–23}, with the possible exception of some parasitoids that attack only after provisioning is complete¹⁶. Progressive provisioning also facilitates the evolution of other social attributes. By prolonging contact between adult and immature wasps, it fosters larval–adult communication and nutritional interdependence^{8,29}. It allows the provision of liquefied and divided prey items directly to larvae^{8,25} and, in larger societies, gives helpers greater control over colony investment decisions³⁰. Progressive provisioning may thus influence not only whether helping evolves, but also the eventual form taken by a society. □

Methods

Work was conducted in north Norfolk (*A. sabulosa*) and at Thursley Common, Surrey (*A. pubescens*). In experiment 1, we placed egg-bearing caterpillars from *Ammophila* field nests in natural-sized laboratory cells hollowed out in sand and covered with glass slides. Under these conditions, we obtained 100% success in rearing 10 unmanipulated offspring of each species through to the cocoon stage. Immature wasps were challenged by placing two *Metopia* maggots, freshly dissected from adult flies, 1 cm away on the caterpillar and then maintaining cells in darkness at room temperature (17–22 °C). Immature wasps were challenged at a range of ages, but eggs were mainly at 7–14 h after oviposition, and larvae within 48 h of hatching. There was no evidence that larval immunity was dependent on age. In experiment 2, control caterpillars were removed from nests and then replaced unmanipulated. In experiment 3, tiny maggots were freshly dissected from *Metopia* in the field, whereas larger maggots had been grown for 1–2 days in the laboratory. In experiment 3, nest contents were replaced within 20 min of their removal. In experiments 2–3, female wasps were individually marked and a maximum of one nest per female was used in each treatment.

Received 18 December 2003; accepted 20 February 2004; doi:10.1038/nature02427.

1. Queller, D. C. in *Natural History and Evolution of Paper-wasps* (eds Turillazzi, S. & West-Eberhard, M. J.) 218–234 (Oxford Univ. Press, Oxford, 1996).
2. Queller, D. C. Extended parental care and the origin of eusociality. *Proc. R. Soc. Lond. B* **256**, 105–111 (1994).
3. Gadagkar, R. Evolution of eusociality: the advantage of assured fitness returns. *Phil. Trans. R. Soc. Lond. B* **329**, 17–25 (1990).
4. Reeve, H. K. in *The Social Biology of Wasps* (eds Ross, K. G. & Matthews, R. W.) 99–148 (Cornell Univ. Press, Ithaca, 1991).
5. Field, J., Shreeves, G., Sumner, S. & Casiraghi, M. Insurance-based advantage to helpers in a tropical hover wasp. *Nature* **404**, 869–871 (2000).
6. Shreeves, G., Cant, M. A., Bolton, A. & Field, J. Insurance-based advantages for subordinate co-foundresses in a temperate paper wasp. *Proc. R. Soc. Lond. B* **270**, 1617–1622 (2003).
7. Bohart, R. M. & Menke, A. S. *Sphecidae Wasps of the World* (Univ. California Press, Berkeley, 1976).
8. Hunt, J. H. Trait mapping and salience in the evolution of eusocial vespids wasps. *Evolution* **53**, 225–237 (1999).
9. Baerends, G. P. Fortpflanzungsverhalten und Orientierung der Grabwespe *Ammophila campestris* Jur. *Tijdschr. Entomol.* **84**, 68–275 (1941).
10. Samuel, C. T. *Factors Affecting Colony Size in the Stenogastrine Wasp Liostenogaster flavolineata*. Thesis, Univ. Malaya (1987).
11. Fabre, J.-H. *The Hunting Wasps* (Hodder & Stoughton, London, 1916).
12. Evans, H. E. Observations on the nesting behaviour of digger wasps of the genus *Ammophila*. *Am. Midl. Nat.* **62**, 449–473 (1959).
13. Field, J. Intraspecific parasitism and nesting success in the solitary wasp *Ammophila sabulosa*. *Behaviour* **110**, 23–46 (1989).
14. Evans, H. E. *The Comparative Ethology and Evolution of the Sand Wasps* (Harvard Univ. Press, Cambridge, Massachusetts, 1966).
15. Wcislo, W. T. Parasitism rates in relation to nest site in bees and wasps (Hymenoptera: Apoidea). *J. Insect Behav.* **9**, 643–656 (1996).
16. O'Neill, K. M. *Solitary Wasps* (Cornell Univ. Press, Ithaca, 2001).
17. Rosenheim, J. A. Nesting behavior and bionomics of a solitary ground-nesting wasp, *Ammophila dysmica* (Hymenoptera: Sphecidae): influence of parasite pressure. *Ann. Entomol. Soc. Am.* **80**, 739–749 (1987).
18. Hager, B. J. & Kuczewski, F. E. Cleptoparasitism of *Ammophila harti* (Fernald) (Hymenoptera: Sphecidae) by *Senotainia vilgians* Allen, with observations on *Phrosinella aurifacies* Downes (Diptera: Sarcophagidae). *Psyche (Stuttg.)* **92**, 7451–7462 (1985).
19. Field, J. Alternative nesting tactics in a solitary wasp. *Behaviour* **110**, 219–243 (1989).
20. Trumbo, S. T. in *Parental Care: Evolution, Mechanisms and Adaptive Significance* (eds Rosenblatt, J. S. & Snowdon, C. T.) 3–51 (Academic, London, 1996).
21. Wyatt, T. D. & Foster, W. A. Parental care in the subsocial intertidal beetle *Bledius spectabilis*, in relation to parasitism by the ichneumonid wasp, *Barycnemis blediator*. *Behaviour* **110**, 76–92 (1989).
22. Godfray, H. C. J. *Parasitoids* (Princeton Univ. Press, New Jersey, 1994).
23. Wyatt, T. D. & Foster, W. A. Leaving home: predation and the dispersal of larvae from the maternal burrow of *Bledius spectabilis*, a subsocial intertidal beetle. *Anim. Behav.* **38**, 778–785 (1989).
24. Queller, D. C. & Strassmann, J. E. Kin selection and social insects. *Bioscience* **48**, 165–175 (1998).
25. Cowan, D. P. in *The Social Biology of Wasps* (eds Ross, K. G. & Matthews, R. W.) 33–73 (Cornell Univ. Press, Ithaca, 1991).
26. Michener, C. D. *The Bees of the World* (John Hopkins Univ. Press, Baltimore, 2000).
27. Schwarz, M. P., Bull, N. J. & Cooper, S. J. B. Molecular phylogenetics of allodapine bees, with implications for the evolution of sociality and progressive rearing. *Syst. Biol.* **52**, 1–14 (2003).
28. Carpenter, J. M. in *The Social Biology of Wasps* (eds Ross, K. G. & Matthews, R. W.) 7–32 (Cornell Univ. Press, Ithaca, 1991).
29. Hunt, J. H. in *Nourishment and Evolution in Insect Societies* (eds Hunt, J. H. & Nalepa, C. A.) 211–244 (Westview, Boulder, 1994).
30. Beekman, M., Komdeur, J. & Ratnieks, F. L. W. Reproductive conflicts in social animals: who has power? *Trends Ecol. Evol.* **18**, 277–282 (2003).

Acknowledgements We thank E. Almond, M. Cant, C. Bridge, A. Cronin, W. Foster and G. Shreeves for advice and comments, and English Nature for access to the study sites. S.B. was funded by a Nuffield Foundation undergraduate research bursary.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.F. (jeremy.field@ucl.ac.uk).

Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10D

Motomichi Matsuzaki¹, Osami Misumi², Tadasu Shin-i³, Shinichiro Maruyama⁴, Manabu Takahara⁵, Shin-ya Miyagishima², Toshiyuki Mori², Keiji Nishida⁶, Fumi Yagisawa⁶, Keishin Nishida², Yamato Yoshida², Yoshiki Nishimura⁶, Shunsuke Nakao⁶, Tamaki Kobayashi⁶, Yu Momoyama⁶, Tetsuya Higashiyama⁶, Ayumi Minoda⁴, Masako Sano³, Hisayo Nomoto³, Kazuko Oishi³, Hiroko Hayashi¹, Fumiko Ohta³, Satoko Nishizaka³, Shinobu Haga³, Sachiko Miura³, Tomomi Morishita³, Yukihiro Kabeya⁷, Kimihiro Terasawa⁷, Yutaka Suzuki⁸, Yasuyuki Ishii⁹, Shuichi Asakawa⁹, Hiroyoshi Takano¹⁰, Niji Ohta⁷, Haruko Kuroiwa², Kan Tanaka⁴, Nobuyoshi Shimizu⁹, Sumio Sugano⁸, Naoki Sato⁷, Hisayoshi Nozaki⁶, Naotake Ogasawara¹¹, Yuji Kohara³ & Tsuneyoshi Kuroiwa²

¹Department of Biomedical Chemistry, Graduate School of Medicine, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

²Department of Life Science, College of Science, Rikkyo (St Paul's) University, 3-34-1 Nishiikebukuro, Toshima-ku, Tokyo 171-8501, Japan

³Center for Genetic Resource Information, National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411-8540, Japan

⁴Institute of Molecular and Cellular Biosciences, University of Tokyo, 1-1-1 Yayoi, Bunkyo-ku, Tokyo 113-0032, Japan

⁵National Institute of Livestock and Grassland Science, 2 Ikenodai, Tsukuba, Ibaraki 305-0901, Japan

⁶Department of Biological Sciences, Graduate School of Science, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

⁷Department of Biochemistry and Molecular Biology, Faculty of Science, Saitama University, 255 Shimo-Ohkubo, Sakura-ku, Saitama, Saitama 338-8570, Japan

⁸Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan

⁹Department of Molecular Biology, Keio University School of Medicine, 35 Shinanomachi, Shinjyuku-ku, Tokyo 160-8582, Japan

¹⁰Department of Biological Science, Graduate School of Science and Technology, Kumamoto University, 2-39-1 Kurokami, Kumamoto 860-8555, Japan

¹¹Graduate School of Biological Science, Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara 630-0101, Japan

Small, compact genomes of ultrasmall unicellular algae provide information on the basic and essential genes that support the lives of photosynthetic eukaryotes, including higher plants^{1,2}. Here we report the 16,520,305-base-pair sequence of the 20 chromosomes of the unicellular red alga *Cyanidioschyzon merolae* 10D as the first complete algal genome. We identified 5,331 genes in total, of which at least 86.3% were expressed. Unique characteristics of this genomic structure include: a lack of introns in all but 26 genes; only three copies of ribosomal DNA units that maintain the nucleolus; and two dynamin genes that are involved only in the division of mitochondria and plastids. The conserved mosaic origin of Calvin cycle enzymes in this red alga and in green plants supports the hypothesis of the existence of single primary plastid endosymbiosis. The lack of a myosin gene, in addition to the unexpressed actin gene, suggests a simpler system of cytokinesis. These results indicate that the *C. merolae* genome provides a model system with a simple gene composition for studying the origin, evolution and fundamental mechanisms of eukaryotic cells.

C. merolae is a small (2 µm diameter) unicellular organism that inhabits sulphate-rich hot springs (pH 1.5, 45 °C) (Fig. 1). The cells of *C. merolae* offer unique advantages for studies of mitochondrial and plastid (chloroplast) divisions¹⁻⁴ because they do not have a rigid cell wall and contain just one nucleus, one mitochondrion and one plastid, divisions of which can be highly synchronized by light/dark cycles⁵. This alga also has the smallest genome of all photo-

synthetic eukaryotes, and contains a minimal set of small membrane-bounded compartments; for example, a microbody (peroxisome), a single Golgi apparatus with two cisternae, coated vesicles, a single endoplasmic reticulum, and a few lysosome-like structures, as well as a small volume of cytosol (Fig. 1)⁶. One of the main points of interest that we discuss regarding this alga focuses on the origin, evolution and fundamental traits (for example, multiplication and differentiation) of single- as well as double-membrane-bounded organelles in plant cells. *C. merolae*, with its complete genomic information, provides an excellent opportunity for addressing such basic questions using microarray and proteome analyses. In addition, from an evolutionary perspective, *C. merolae* has other noteworthy properties that allow us to study the origin of eukaryotic cells, primary endosymbiosis between cyanobacteria and eukaryotic hosts, and secondary endosymbiosis between red algae and their hosts.

Samples of *C. merolae* 10D were isolated⁷ from the hot spring algal collection provided by G. Pinto (Naples University). The entire *C. merolae* genome was sequenced using the random sequencing method (see Methods). We obtained 16,520,305 base pairs (bp; approximately 99.98% of the estimated total length) of the nuclear genome sequence (Fig. 2, Table 1, and Supplementary Fig. 1 and Supplementary Table 1) with 46 gaps. The genome is distributed among the 20 chromosomes and ranges in size from approximately 0.42 to 1.62 Mb. No significant deviation in statistical parameters, such as base composition and gene density, were observed among the chromosomes (Supplementary Table 1). The overall G+C composition was 55.0%. The dinucleotide CpG in the *C. merolae* genome was exceptionally over-represented (1.151) compared with the expected value from observations of G+C content; it is generally underrepresented in other eukaryote genomes (Table 1).

The putative repeat unit of telomeres in *C. merolae* is

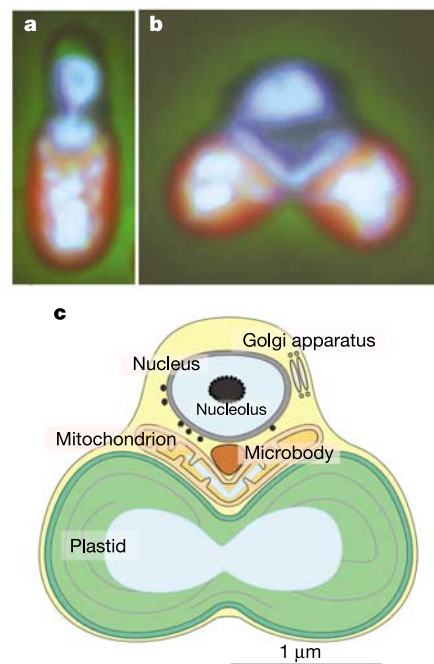


Figure 1 The unicellular red alga *C. merolae* 10D. Phase contrast-fluorescent images of the interphase (a) and dividing (b) cells show localization of nuclear (top), mitochondrial (middle) and plastid DNA (bottom in blue/white) after DAPI staining. The plastids emit red autofluorescence. The schematic dividing cell (c) contains a nucleus, a V-shaped mitochondrion, a dumb-bell-shaped plastid, a microbody and a Golgi apparatus, divisions of which can be highly synchronized by light/dark cycles.

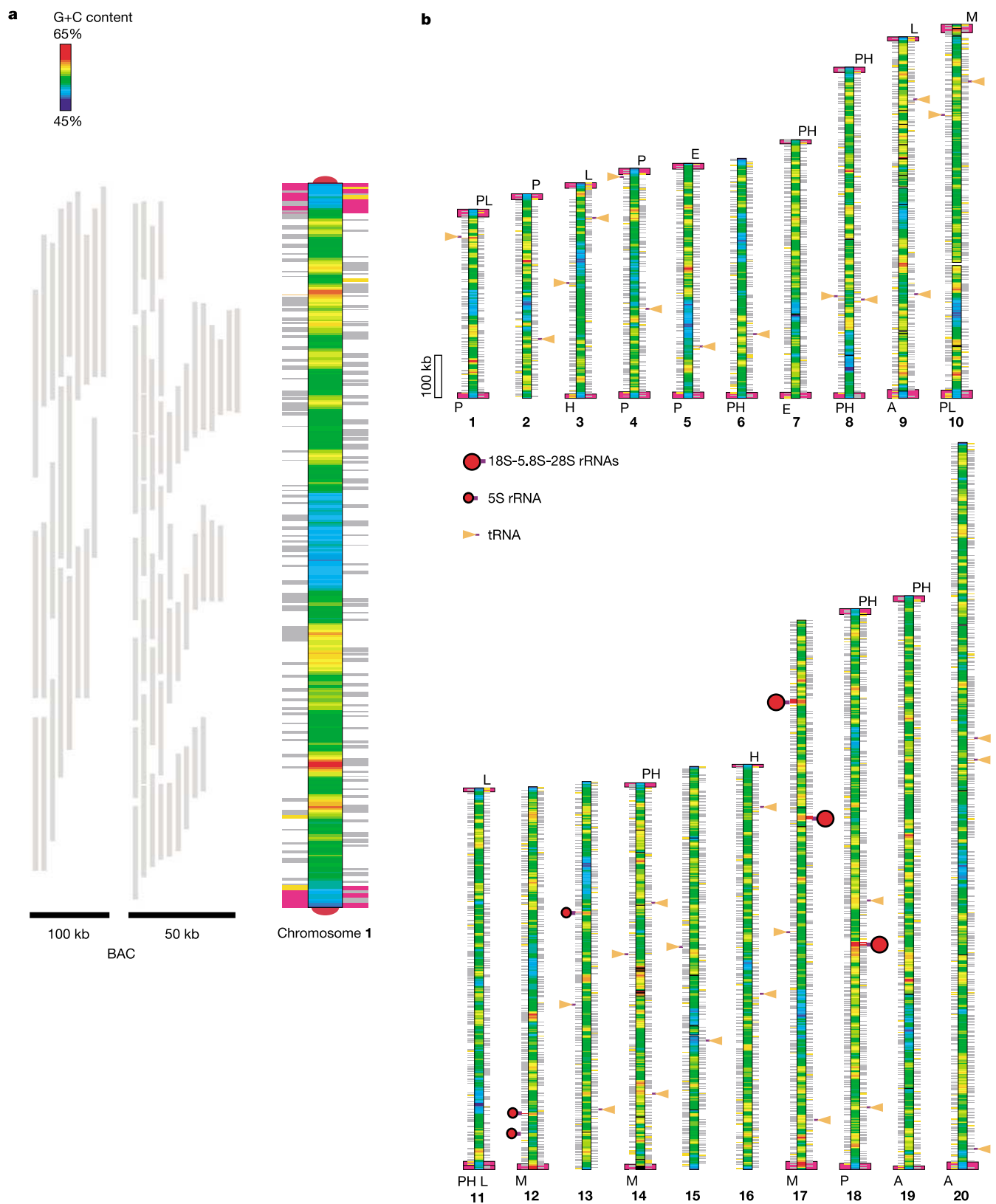


Figure 2 Representation of *C. merolae* chromosomes. **a**, Chromosome 1 and mapping of the BAC clones (50 kb and 100 kb). The chromosome is represented as a bar with pseudo-colour assignments of local G+C contents. Side rods represent genes and gene-like elements (left for transcription towards the top, and right for the bottom). Telomere and subtelomeric elements are indicated respectively as semicircles and purple

rectangles at each end of the chromosome. **b**, A bird's-eye view of 20 chromosomes showing G+C contents, genes, subtelomeric elements (designations on the side: P, H, L, A, E, and so on), and RNA genes. A putative centromeric (A+T-rich) region is located on each chromosome.

Table 1 Nuclear and organellar genomes of *C. merolae* and comparison to other genomes

Feature	<i>C. merolae</i>	<i>S. pombe</i>	<i>S. cerevisiae</i>	<i>A. thaliana</i>	<i>P. falciparum</i>
Nucleus					
No. of chromosomes	20	3	16	5	14
Sequenced length (bp)	16,520,305	12,462,637*	12,495,682*	115,409,949*	22,853,764
G+C content (%)	55.0	36.0	38.3	34.9	19.4
CpG occurrence (Obs./Exp.)	1.151	0.886	0.803	0.724	0.765
No. of genes	5,331	4,929	5,770	25,498	5,268
Mean gene length† (bp)	1,552	1,426	1,424	1,310	2,283
Gene density (bp per gene)	3,099	2,528	2,088	4,526	4,338
Per cent coding	44.9	57.5	70.5	28.8	52.6
Genes with introns (%)	0.5	43.0	5.0	79.0	53.9
No. of introns	27	4,730	272	107,784	7,406
Mean length of intron (bp)	248	81	NA	170	179
Mean length of exons (bp)	1,540	ND	ND	170	949
No. of tRNA genes	30	174	274	620	43
No. of 5S rRNA genes	3	30	100–150	1,000	3
No. of 18S, 5.8S and 28S rRNA units	3	200–400	100–150	700–800	7
Mitochondrion					
Genome size (bp)	32,211	19,431	85,779	366,924	5,967
No. of protein genes	34	10	29	58	3
Density (bp per protein genes)	947	1,943	2,958	6,326	1,989
Plastid					
Genome size (bp)	149,987	NA	NA	154,478	29,422
No. of protein genes	208	NA	NA	79	30
Density (bp per protein genes)	721	NA	NA	1,955	981
Genes with introns (%)	0	NA	NA	18	0

Data for the other organisms are from ref. 31. ND, not determined; NA, not applicable.

*Ribosomal DNA repeats were not sequenced.

†Excluding introns.

GGGGGAAT, and as far as could be determined experimentally, it is found on both ends of the chromosomes. In addition, several sequence elements up to 20 kilobases (kb) in length were duplicated in 30 of the 40 putative subtelomeric regions (Fig. 2, Supplementary Fig. 1). Each chromosome has, in varying degrees, a single A+T-rich region on its mid-section. As chromosomal centromeric regions generally have a biased base composition, this A+T-rich region possibly defines centromeres (Fig. 2). The centromeres were confirmed via immunological experiments using antibodies against CENP-A, which was identified in the *C. merolae* genome (data not shown). Unlike many other eukaryotes, the *C. merolae* genome does not contain tandem repeated arrays of ribosomal RNA (rRNA) genes (Fig. 2). A single rRNA gene unit (18S-5.8S-28S) was discovered on three separate loci. The three units were virtually identical in sequence. Moreover, *C. merolae* has only three copies of the 5S rRNA gene, the sequences of which are also almost identical.

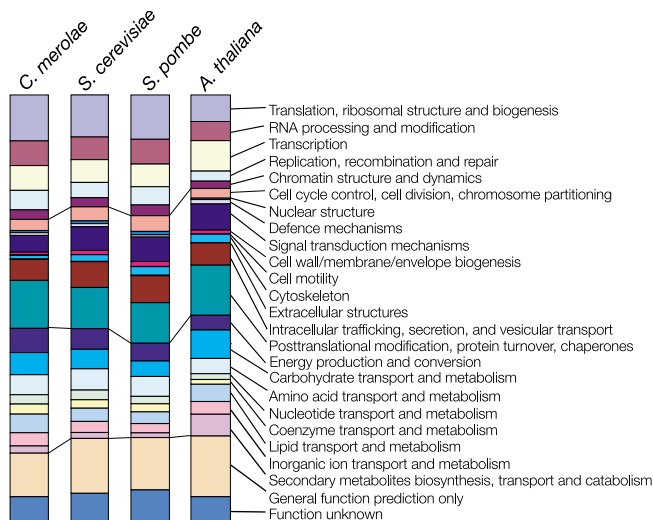


Figure 3 Comparison of the functional classification of *C. merolae* proteins with other organisms. Columns represent the proportion of proteins assigned to KOG classification of each organism: *C. merolae*, *S. cerevisiae*, *S. pombe* and *A. thaliana* in a left-to-right fashion. The actual numbers of proteins assigned to each classification are given in Supplementary Table 3.

Therefore, *C. merolae* has the smallest set of rRNA genes among all eukaryotes thus far studied. These results might be related to the existence of a single small nucleolus without nucleolus-associated chromatin. Furthermore, they also promote studies on the origin and formation of the nucleolus, because even prokaryotic cells with more than three copies of rRNA gene units do not have a nucleolus.

A full-length complementary DNA (cDNA) library was used to map expressed genes within the *C. merolae* genome. Fortunately, 99.85% of the expressed sequence tags (ESTs) were mapped on the genome sequence. In addition, many cDNA clones encoded a single open reading frame (ORF) bridging both end sequences. This suggests that most *C. merolae* genes lack introns. The predicted genes were automatically annotated using several databases (see Methods and Supplementary Information). As a result, 5,331 genes were identified, and 86.3% of them had corresponding ESTs (Supplementary Table 2). The number of genes in the *C. merolae* genome is similar to those found in yeasts and malarial parasites, despite the great ecological differences between these species (Table 1). Furthermore, the genes of the *C. merolae* genome are remarkable for their paucity of introns. Only 26 genes (0.5% of the protein genes) contained introns, and all but one of them had only a single intron. These introns had strict consensus sequences (Supplementary Fig. 2 and Supplementary Table 3).

Figure 3 summarizes the repertoire of *C. merolae* proteins on the basis of their assignment to eukaryotic clusters of orthologous groups (KOGs)⁸. Of the 4,771 predicted proteins, 2,536 were assigned to KOGs, by emulating the NCBI KOGnitor service (<http://www.ncbi.nlm.nih.gov/COG/new/kognitor.html>). The distribution of the functional classification of *C. merolae* was compared with those of other free-living unicellular eukaryotes, such as *Saccharomyces cerevisiae*⁹ and *Schizosaccharomyces pombe*¹⁰, and a higher plant *Arabidopsis thaliana*¹¹. The distribution was on the whole similar to both yeasts which have similar genome size although *C. merolae* cells contain plastids. The lowered proportion of genes for 'secondary metabolites biosynthesis, transport and catabolism' found in these unicellular organisms, as compared with that of *A. thaliana*, might reflect their simple cellular organizations (Fig. 3, Supplementary Table 4).

In *C. merolae*, the division of double-membrane-bounded mitochondria and plastids involves a dynamic trio: an FtsZ ring of bacterial origin, electron dense mitochondrial/plastid dividing rings (MD and PD rings), and eukaryotic mechanochemical dynamin

rings. Four genes representing mitochondrial FtsZ (*FtsZ2-1* and *FtsZ2-2*) and plastid FtsZ (*FtsZ1-1* and *FtsZ1-2*) were identified^{12,13}. A large gene family consisting of more than 10 members encoding functionally diverse dynamins with a wide range of membrane pinching roles have been found in other organisms; however, only two dynamin genes (*C. merolae* *Dnm1* and *Dnm2*) are found, with a role in the later stages of the mitochondrion¹⁴ and plastid¹³ division, respectively. These findings suggest that plastids and mitochondria divide in a similar way, using very common systems consisting of the amalgamation of bacterial and eukaryotic rings. The dynamic trio of plastid division is conserved in lower algae to higher plants^{15,16}. With mitochondrial divisions, however, whilst dynamin rings are retained in higher organisms, FtsZ and MD rings are not clearly observed, and it is possible that they were replaced by other systems during eukaryotic evolution¹³. MD/PD ring genes are yet unknown, although their identification should be accelerated by works such as this. Although the microbody, a single-membrane-bounded organelle, divides by binary fission in *C. merolae*¹⁷, it lacks Pex11p, which is a known key regulator of microbody division and proliferation¹⁸.

The following proteins related to cell motility and cytokinesis were encoded in *C. merolae* (Supplementary Table 5); one set of tubulin, two actins, five proteins of the kinesin family, and several intermediate filament proteins. However, no genes encoding myosin or proteins containing dynein motor domains were found. The absence of the myosin gene is consistent with the fact that electron microscopy and immuno-detection¹⁹ techniques did not detect microfilaments of actin; cDNA clones for actin genes were also not obtained. In the red alga *Cyanidium caldarium* RK-1, which is closely related to *C. merolae* but has a genome double the size⁷, cells divide using a contractile ring of actin filaments¹⁹. *C. merolae* cells therefore seem to divide using a system that is simpler than that of actomyosin.

C. merolae has noteworthy properties, which are relevant for examining the origin of eukaryotes, and primary and secondary plastid endosymbiosis²⁰. Only 30 transfer RNAs (tRNAs) were detected in the nuclear genome using the program tRNAscan-SE with relaxed parameter settings (Fig. 2). Some of these tRNA genes showed possible archaeal features, namely, ectopic introns and anticodon GAU for tRNA-Ile (Supplementary Fig. 3). Four of these tRNA genes seemed to have introns in the D-loop region, whereas the introns of eukaryotic tRNA genes are limited to a site 3' to the anticodon. As ectopic tRNA introns have been reported in some archaeal genomes²¹, this could explain the paucity of detected tRNA genes in *C. merolae*; tRNAscan-SE might have overlooked other tRNAs owing to the existence of unknown types of ectopic introns. Another point to note is that *C. merolae* possesses a single tRNA-Ile with anticodon GAU, which has not been observed in eukaryotes, but only in prokaryotes²¹.

Standard sets of photosystem genes, including those encoding phycobilisome components, were observed in *C. merolae*. Many of them (11 PSI genes and 17 PSII genes) are encoded in the plastid genome²², while PsbO, P, U, Z as well as a distant PsbQ homologue are encoded in the nuclear genome. Although only PsbU and PsbZ were previously identified in red algal PSII, the localization of PsbP and putative PsbQ in PSII, as recently suggested in *Synechocystis* sp. PCC6803²³ is an interesting subject of proteomic study. The genes *psaH*, *N*, *X*, as well as *psbS* and *ndh* genes are not encoded in either the plastid or the nuclear genomes. Therefore, the photosystems of *C. merolae* lack various mechanisms for dissipating excessive light energy.

Enzymes of the Calvin cycle in plants are known to be a mosaic of enzymes originating from cyanobacteria-like ancestors of an endosymbiont and its eukaryotic host²⁴. Red algal ribulose-1,5-bisphosphate carboxylase/oxygenase (Rubisco) is known to be a product of horizontal gene transfer²⁵. The origin of other Calvin cycle enzymes is essentially identical in *C. merolae* and *A. thaliana* (Supplementary Fig. 4 and Supplementary Table 6). It is highly probable that the complex and mosaic origin of Calvin cycle

enzymes derived from common ancestors of green plants and red algae, and no essential changes occurred after the separation of the two lineages. This is strong support for the concept of a single event of primary plastid endosymbiosis. Among the known translocon proteins of plastids, Toc34, Toc75, Tic20, Tic22 and Tic110 were encoded in the *C. merolae* genome, but other proteins such as Toc159, Tic40 and Tic55 were not found. Results of phylogenetic analysis of the five translocon components (to be published elsewhere) also suggest the concept²⁶.

Another aspect of the comparative genomics of the red algal genome is secondary endosymbiosis. Cryptophytes are thought to retain a remnant of the endosymbiotic red algal nucleus, the nucleomorph, in the periplastidic compartment. The sequencing of the cryptophyte alga *Guillardia theta* nucleomorph genome revealed a number of curious architectural features that might be shared by the genome of red algae²⁷. *C. merolae* chromosomes showed multiple subtelomeric duplications, but did not contain rRNA gene clusters such as those of the nucleomorph genome. This implies that the telomeric rRNA gene clusters observed in the nucleomorph genome, as well as other prominent genome structures such as overlapped genes, appeared after secondary symbiosis. It is also notable that ectopic tRNA introns are also reported in nucleomorph tRNAs²⁸. Details of the comparisons with the nucleomorph genome will be presented elsewhere.

Light signal transduction is critical for the growth and differentiation of photoautotrophic organisms. As the division of *C. merolae* cells is synchronized by light, an elaborate mechanism for light signal transduction must exist. Several putative blue light receptor (cryptochrome) genes were found in *C. merolae*, whereas no genes encoding phytochromes and phototropins were identified. As bacterial phytochrome genes are only found in some species of cyanobacteria with large genomes²⁹, the ancestor of plastids might be an ancestral cyanobacterium without phytochromes. This also suggests that the phytochromes of higher plants might not be of cyanobacterial origin. In higher plants, various signalling pathways (such as the two-component system consisting of histidine kinases and response regulators as well as a MAP kinase cascade) are involved in the signal transduction of various hormones, and in the development of organs. In *C. merolae*, the presence of only a single candidate for histidine kinase and a dozen MAP kinase-related molecules is suggested. However, there are no response regulators other than those that are plastid-encoded, trimeric G protein and adenylate cyclase. Thus, *C. merolae* appear to use only a limited repertoire of signal transduction mechanisms, which corroborates the lack of cell differentiation in this alga.

C. merolae is an alga in which all of the three genome compartments—nucleus, mitochondrion (32,211 bp)³⁰ and plastid (149,987 bp)²²—have been sequenced. Such information is a prerequisite for future studies on proteomics, expression analysis using microarrays, and structural biology with heat-stable proteins that are unique among eukaryotes. All of this information will, in turn, help elucidate the origin, evolution and fundamental mechanisms of the single- as well as double-membrane-bounded organelles, and ultimately all photosynthetic eukaryotes. In addition, this hot spring alga will be useful in analysing the mechanisms of heat and acid tolerance in eukaryotic cells. □

Methods

Whole genome shotgun sequencing

We sequenced the *C. merolae* genome by the whole genome random sequencing method (see Supplementary Information for details). About 335,000 insert ends were sequenced, which covered the genome 11 times. BAC libraries with two subsets were constructed and a large-scale full-length cDNA library from cells cultured under various growth conditions prepared. The sequences were assembled using Phrap, further examined by referring to another assembly using ARACHNE, and edited using CONSED. The scaffolds were built within the hybridization groups using read-pair information from the BAC, shotgun and cDNA clones. The gaps between the contigs were closed by primer walking PCR, and mate-pair clone and BAC clone sequencings.

Gene identification and annotation

We principally used two strategies for gene prediction and combined the results (see Supplementary Information for details). (1) Each read-pair of cDNA clones was mapped on the contigs using BLAST and putatively transcribed regions were determined by clustering the mapped pairs. (2) ORFs likely to encode a protein showing similarity to known proteins or having known motifs were identified respectively by the BLASTP program with GenBank nr database, or a HMMER program with a Pfam database. A functional classification was performed based on the NCBI KOG. The tRNA genes were detected using the tRNAscan-SE program with relaxed parameters (-X 15 -I -36).

Received 13 November 2003; accepted 5 February 2004; doi:10.1038/nature02398.

1. Kuroiwa, T. The primitive red algae: *Cyanidium caldarium* and *Cyanidioschyzon merolae* as model system for investigating the dividing apparatus of mitochondria and plastids. *Bioessays* **20**, 344–354 (1998).
2. Kuroiwa, T. *et al.* The division apparatus of plastids and mitochondria. *Int. Rev. Cytol.* **181**, 1–41 (1998).
3. McFadden, G. I. & Ralph, S. A. Dynamin: the endosymbiosis ring of power. *Proc. Natl Acad. Sci. USA* **100**, 3557–3559 (2003).
4. Surridge, D. Ancient rings. *Nature* **422**, 275 (2003).
5. Terui, S., Suzuki, K., Takahashi, H., Itoh, R. & Kuroiwa, T. High synchronization of chloroplast division in the ultramicro-alga *Cyanidioschyzon merolae* by treatment with both light and aphidicolin. *J. Phycol.* **31**, 958–961 (1995).
6. Kuroiwa, T. *et al.* Comparison of ultrastructures between the ultra-small eukaryote *Cyanidioschyzon merolae* and *Cyanidium caldarium*. *Cytologia (Tokyo)* **59**, 149–158 (1994).
7. Toda, K., Takahashi, H., Itoh, R. & Kuroiwa, T. DNA contents of cell nuclei in two Cyanidiophyceae: *Cyanidioschyzon merolae* and *Cyanidium caldarium* Forma A. *Cytologia (Tokyo)* **60**, 183–188 (1995).
8. Tatusov, R. L. *et al.* The COG database: an updated version includes eukaryotes. *BMC Bioinform.* **4**, 41 (2003).
9. Goffeau, A. *et al.* Life with 6000 genes. *Science* **274**, 546, 563–567 (1996).
10. Wood, V. *et al.* The genome sequence of *Schizosaccharomyces pombe*. *Nature* **415**, 871–880 (2002).
11. Arabidopsis Genome Initiative, Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815 (2000).
12. Takahara, M. *et al.* A putative mitochondrial *ftsZ* gene is present in the unicellular primitive red alga *Cyanidioschyzon merolae*. *Mol. Gen. Genet.* **264**, 452–460 (2000).
13. Miyagishima, S., Nishida, K. & Kuroiwa, T. An evolutionary puzzle: chloroplast and mitochondrial division rings. *Trends Plant Sci.* **8**, 432–438 (2003).
14. Nishida, K. *et al.* Dynamic recruitment of dynamin for final mitochondrial severance in a primitive red alga. *Proc. Natl Acad. Sci. USA* **100**, 2146–2151 (2003).
15. Kuroiwa, H., Mori, T., Takahara, M., Miyagishima, S. & Kuroiwa, T. Chloroplast division machinery as revealed by immunofluorescence and electron microscopy. *Planta* **215**, 185–190 (2002).
16. Gao, H., Kadrijan-Kalbach, D., Froehlich, J. E. & Osteryoung, K. W. ARC5, a cytosolic dynamin-like protein from plants, is part of the chloroplast division machinery. *Proc. Natl Acad. Sci. USA* **100**, 4328–4333 (2003).
17. Miyagishima, S. *et al.* Microbody proliferation and segregation cycle in the single-microbody alga *Cyanidioschyzon merolae*. *Planta* **208**, 326–336 (1999).
18. Marshall, P. A. *et al.* Pmp27 promotes peroxisomal proliferation. *J. Cell Biol.* **129**, 345–355 (1995).
19. Takahashi, H. *et al.* A possible role of actin dots in the formation of the contractile ring in the ultramicro alga *Cyanidium caldarium* RK-1. *Protoplasma* **201**, 115–119 (1998).
20. Nozaki, H. *et al.* The phylogenetic position of red algae revealed by multiple nuclear genes from mitochondria-containing eukaryotes and an alternative hypothesis on the origin of plastids. *J. Mol. Evol.* **56**, 485–497 (2003).
21. Marck, C. & Grosjean, H. tRNomics: analysis of tRNA genes from 50 genomes of Eukarya, Archaea, and Bacteria reveals anticodon-sparing strategies and domain-specific features. *RNA* **8**, 1189–1232 (2002).
22. Ohta, N. *et al.* Complete sequence and analysis of the plastid genome of the unicellular red alga *Cyanidioschyzon merolae*. *DNA Res.* **10**, 67–77 (2003).
23. Kashino, Y. *et al.* Proteomic analysis of a highly active photosystem II preparation from the cyanobacterium *Synechocystis* sp. PCC 6803 reveals the presence of novel polypeptides. *Biochemistry* **41**, 8004–8012 (2002).
24. Martin, W. & Schnarrenberger, C. The evolution of the Calvin cycle from prokaryotic to eukaryotic chromosomes: a case study of functional redundancy in ancient pathways through endosymbiosis. *Curr. Genet.* **32**, 1–18 (1997).
25. Ohta, N., Sato, N., Ueda, K. & Kuroiwa, T. Analysis of a plastid gene cluster reveals a close relationship between *Cyanidioschyzon* and *Cyanidium*. *J. Plant Res.* **110**, 235–245 (1997).
26. Cavalier-Smith, T. Genomic reduction and evolution of novel genetic membranes and protein-targeting machinery in eukaryote-eukaryote chimaeras (meta-algae). *Phil. Trans. R. Soc. Lond. B* **358**, 109–134 (2003).
27. Douglas, S. *et al.* The highly reduced genome of an enslaved algal nucleus. *Nature* **40**, 1091–1096 (2001).
28. Zauner, S. *et al.* Chloroplast protein and centrosomal genes, a tRNA intron, and odd telomeres in an unusually compact eukaryotic genome, the cryptomonad nucleomorph. *Proc. Natl Acad. Sci. USA* **97**, 200–205 (2000).
29. Montgomery, B. L. & Lagarias, J. C. Phytochrome ancestry: sensors of bilins and light. *Trends Plant Sci.* **7**, 357–366 (2002).
30. Ohta, N., Sato, N. & Kuroiwa, T. Structure and organization of the mitochondrial genome of the unicellular red alga *Cyanidioschyzon merolae* deduced from the complete nucleotide sequence. *Nucleic Acids Res.* **26**, 5190–5198 (1998).
31. Gardner, M. J. *et al.* Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**, 498–511 (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank many colleagues for discussions, especially K. Kita, Y. Watanabe, H. Fujiwara and T. Q. Ueda. We also thank Trans New Technology, Inc. for providing computational resources. This work was supported by Grants-in-Aid for Scientific Research on Priority Areas “Genome” from the Ministry of Education, Culture, Sports, Science, and Technology of Japan, and a Grant-in-Aid from the Promotion of Basic Research Activities for Innovative Biosciences (ProBRAIN).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to T.K. (tsune@rikkyo.ne.jp) or M.M. (mzaki@biol.s.u-tokyo.ac.jp). Chromosome sequences were submitted to DDBJ with accession numbers AP006483–AP006502 (chromosome 1–20) and AP006600–AP006614 (unassigned contigs). Sequences and annotation are available at <http://merolae.biol.s.u-tokyo.ac.jp/> or <http://dolphin.lab.nig.ac.jp/publish/>.

Long-lasting sensitization to a given colour after visual search

Chia-huei Tseng^{1*}, Joetta L. Gobell^{1*} & George Sperling^{1,2}

¹Department of Cognitive Sciences, and ²Neurobiology and Behavior, and Institute of Mathematical Behavioral Sciences, University of California, Irvine, California 92697, USA

* Present addresses: Center of Cognitive Science, Rutgers University, New Brunswick, New Jersey 08854, USA (C.-h.T.); Department of Psychology and Center for Neural Science, New York University, New York, New York 10003, USA (J.L.G.)

Visual attention enables an observer to select specific visual information for processing. In an ambiguous motion task in which a coloured grating can be perceived as moving in either of two opposite directions depending on the relative salience of two colours in the display, attending to one of the colours influences the direction in which the grating appears to move¹. Here, we use this secondary effect of attention in a motion task to measure the effect of attending to a specific colour in a search task. Observers performed a search task in which they searched for a target letter in a 4 × 4 coloured matrix. Each of the 16 squares within a matrix was assigned one of four colours, and observers knew that the target letter would appear on only one of these colours throughout the experiment. Observers performed the ambiguous motion task before and after the search task. Attending to a particular colour for a brief period in the search task profoundly influenced the perceived direction of motion. This effect lasted for up to one month and in some cases had to be reversed by practising searches for the complementary colour, indicating a much longer-persisting effect of attention than has been observed previously.

To investigate the consequences of attending to a particular colour, we designed a search task that requires observers to attend to a particular colour and to ignore all other colours. The task is to report the location of a target letter among other ‘distracter’ characters. Each trial consists of ten consecutive 4 × 4 matrices displayed rapidly (Fig. 1a). Each of the 16 squares of a matrix is randomly assigned one of four colours (red, green, yellow or blue). The initial frame rate was two matrices per second and it was increased as observers’ performance improved to as high as 19 matrices s⁻¹ for the fastest observer. At the average final speed of 10 frames s⁻¹, each colour display was shown for 50 ms and then the to-be-searched letters and numbers appeared superimposed on the coloured squares for an additional 50 ms. The purpose of advancing the colour matrix relative to the target was to induce the observers to use the colour to find the target. A target colour, either red or green, was assigned to each observer for the duration of the first phase of the experiment.

At the beginning of each trial, observers were shown a randomly

selected letter that was to be the target for that trial. During the search sequence, this letter appeared on the target colour square once and on non-target colour squares an average of 3.6 times. The task was to report the spatial location of the square on which the target occurred. The target letter was to be ignored when it occurred on non-target colour squares. If observers were able to use colour information optimally, they would need to search only one-quarter of the squares for the target. As observers practiced, whenever performance reached 95% correct, the frame rate was increased. One-hour sessions were conducted until observers' performance failed to improve in three consecutive runs, which took 4–7 h. At the average final rate of ten search frames s^{-1} the duration of the search sequence itself was 1 s (ten search frames presented consecutively over a duration of 1 s; that is, 10 frames s^{-1}). In a 1 h session of 300 trials, there would be only 5 min of actual viewing of the stimulus. Throughout the entire experiment, no observer was exposed to the search displays for more than 1 h.

Sensitization to the target colour (either red or green) was measured by a third-order motion paradigm (Fig. 1b)¹. The experiment was conducted in a different room and with a different apparatus than the search experiment. A trial of the motion paradigm is composed of five frames, with a 90° phase shift between consecutive frames. Odd frames contained horizontal isoluminant red–green sine wave gratings (the chromaticities of the red and green in the motion task were quite different from those used in the

search task) (Fig. 1), whereas even frames contained thin stripes of high-contrast noise alternating with thick stripes of low-contrast noise. In both even and odd frames, the average luminance is constant throughout. When five frames are displayed successively, the perceived direction of motion depends on the relative salience of the red and green, as there are no traditional cues (for example, luminance or contrast) to motion perception. When red and green are equally salient, no consistent direction of motion is perceived.

Other things being equal, the salience of a coloured area is proportional to colour difference from the background. On our grey background, greater saturation means greater salience. Before the search task, observers performed the motion task with a range of colour gratings that varied in their red/green saturation ratios. Their responses are represented as a psychometric function depicting the percentage of trials in which motion was seen in a direction consistent with a particular colour (Fig. 2a, b, black line). This function was taken as the baseline. When training on the search task was concluded (after performance no longer improved), another set of motion measurements were made and a new psychometric function was recorded (Fig. 2a, b, coloured line). The lateral shift between 'before' and 'after' psychometric functions is equivalent to an increase in the saturation of the attended colour by an average of

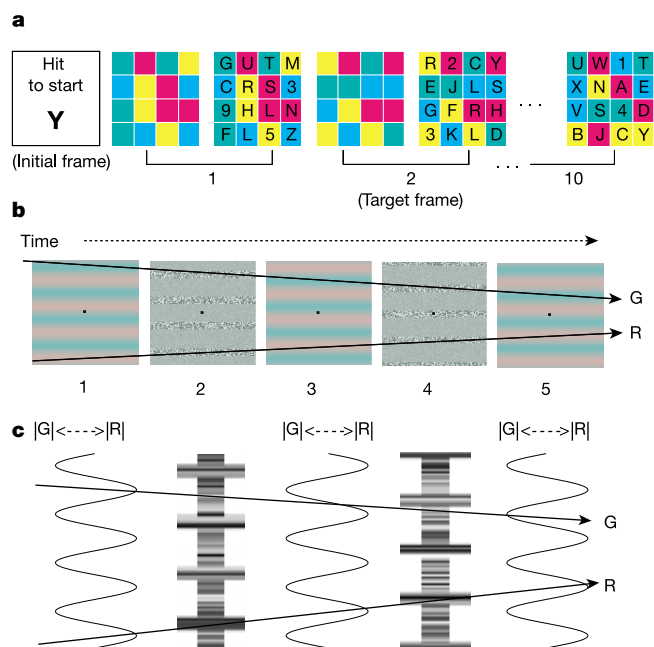


Figure 1 The visual search task and the third-order motion paradigm. **a**, The search task. A target colour (either red or green) is assigned to each observer (red in this example). An initial frame shows the observer the target letter for that particular trial (Y in this example). The search task contains ten search frames in which the coloured background is visible throughout but the letters only for the second half of the interval. The observer's task is to detect the target letter and report its location only when it appears on the target colour. Only one-half of trials contained the target item. In the other half, observers press the space bar to indicate there was no target item detected. **b**, The motion task. The third-order motion task is composed of five frames. The even frames contain a high-contrast thin stripe, which is perceived as figure against a low-contrast background. The odd frames contain a red–green grating modulated along the long–medium wavelength cardinal axis in colour space. Frames are arranged so that apparent motion can be perceived in either of two opposed directions or in neither direction. The perceived direction of apparent motion depends on whether the red (R) or green (G) of that display has higher salience. **c**, Schematic illustration of green–red and black–white contrast in the motion frames.

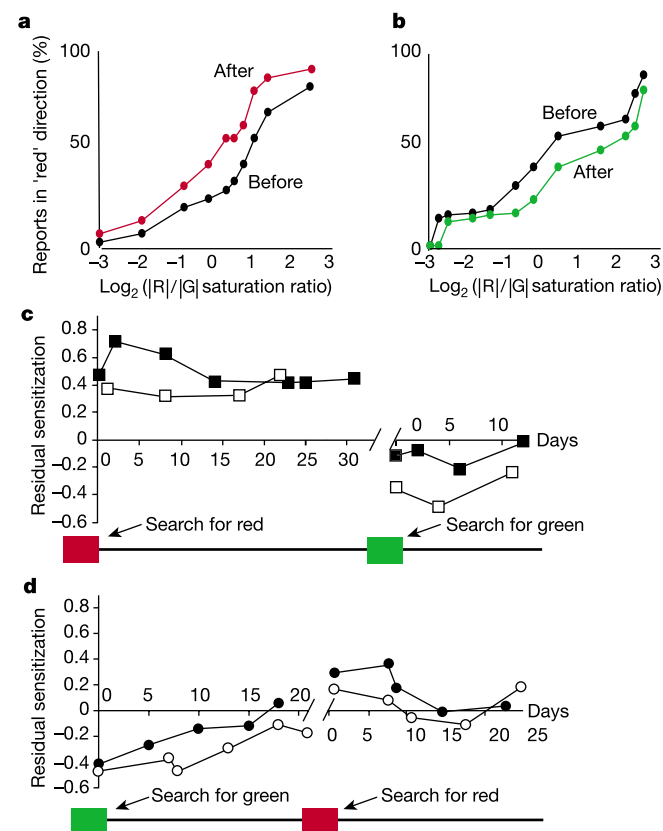


Figure 2 Results. **a**, Typical psychometric data from the third-order ambiguous motion paradigm before (black curve) and after (red curve) the search for red. The percentage of trials in which observers report motion in the red direction is shown as a function of the saturation of red divided by the saturation of green. The horizontal shift indicates that, after the search task, the searched-for colour behaves in the motion task as though it has become more saturated, that is, more salient. Data are for observer JYC. **b**, Typical psychometric data after search for green. Data are for observer CHT. **c**, Data from two observers (CCW, filled symbols; TSAI, open symbols) showing that the residual sensitization to red produced by searching for red lasts for weeks. Positive values indicate sensitization to red and negative values indicate sensitization to green. Positive values are obtained after observers search for red. The effect was reversed after observers searched for green. **d**, Data from observers (TWH, filled symbols; YCT, open symbols) who searched for green first, and then switched to searching for red.

30%. When observers were subsequently trained to search for a different colour, the shift in the psychometric function was reversed, and the observers become sensitized to the new search colour.

The third-order motion paradigm was used to measure how long the sensitization effect persisted. After finishing the search task, observers returned at regular intervals to perform the motion task, producing a new psychometric function at each visit. We take the sum of differences at all points between the new and baseline psychometric functions as an index of the magnitude of the sensitization effect. The data show that there is a long-term sensitization for all of our observers and that this sensitization persists for weeks, or even a month (Fig. 2c, d).

We explain how our third-order motion paradigm reflects observers' sensitization in a search task in terms of the hierarchical motion theory². This theory proposes that the third-order motion system detects movements of salience. Salience is recorded in a salience map $S(x, y, t)$ in which at time t , at location x, y , the areas indicated as 'figure' are marked with positive values whereas locations indicated as 'ground' or as unimportant are marked with zero or negative values. Some areas may be ambiguously classified as figure or ground, and attention can have a role in resolving this ambiguity. As a consequence of performing the search task, the relative importance of colours changes because one of them is the target colour and the others are distracter colours. This change in importance, which is understood at a high cognitive level, is used to influence how locations containing those attended and unattended colours are recorded in the salience map. The third-order motion system is assumed to use the salience values to compute motion direction just like the luminance motion system uses luminance values. By recording the performance change in the motion task, we are able to measure quantitatively the sensitization to colour produced by a completely different task.

It is surprising that the colour sensitization persists for weeks. The lifetime of most simple after-effects is measured in seconds. The most notable persisting after-effect is the McCollough effect³ in which, after viewing a coloured grating for seconds or minutes, black-white stripes of a similar slant and spatial frequency appear to have a colour tint opposite to the adapting grating. The McCollough effect³ (and several similar after-effects⁴⁻⁷) typically lasts for seconds or minutes, although subsequent studies have demonstrated that it can be made to last longer when the adapting procedure is very much longer⁷⁻¹³. Additionally, viewing through a coloured filter for several days can produce a weeks-long perceptual colour after-effect¹⁴. Again, the observer becomes relatively less sensitive to the colour passed by the filter, a kind of perceptual colour renormalization.

Previously reported long-lasting after-effects, such as the McCollough⁸, Neitz¹⁴ and others⁷⁻¹³, differ from the present attentional effect in several important ways. (1) These after-effects are opposite to the original stimulus, for example, exposure to red desensitizes red relative to green, exactly the opposite of our observed sensitization; (2) the initial adaptation stimuli and subsequent test stimuli are quite similar, whereas our search adapting and motion testing procedure differ in task, spatial structure and precise colour coordinates; (3) not surprisingly therefore, the typical test procedure, being similar to the adapting procedure, interferes with the persistence of the after-effect⁸. Our testing procedure (the third-order motion task) involved a great many motion trials on each test occasion but seemed to produce no diminution of attentional sensitization for at least some of the subjects for at least during the first month of repeated tests.

In perceptual learning an observer performs a discrimination task with a very specific stimulus. Over the course of many trials, and with interposed intervals of sleep, performance can improve markedly¹⁵; however, enhanced performance is restricted to the specific location in the visual field where the stimulus normally appears¹⁶, to the specific stimulus itself¹⁷⁻²⁰, it frequently fails to generalize to the unexposed eye^{21,22}, and it can persist for years¹⁵. In all of these

respects, it differs from the attention-produced enhanced salience of a specific colour observed here. In our study, observers' sensitization to the target colour can transfer between two unrelated tasks, demonstrating that the sensitization must be registered in a common area that both tasks can access. We propose a salience map to conceptualize the dynamic process that instantiates the sensitization to colour.

The third-order motion display is a good measure of attention to colour whether attention is manipulated directly by instruction or by other methods, such as by a search task. We found that searching for less than 1 h (distributed over 4-7 h) increased the salience of the searched-for colour by an amount that was equivalent to making the colour 1.3 times more saturated. The sensitization persisted for weeks. It is a long-term perceptual modification that is very different from adaptation, contingent after-effects and perceptual learning. □

Methods

Visual search task

Observers reported the location of a target letter among other distracters in ten consecutive 4×4 matrices displayed rapidly (Fig. 1a). The target letter varied in each trial, and observers were informed of its identity before search. A target colour, either red or green, was assigned to each observer for the duration of the search phase of the experiment. If present, the target letter occurred only once on the target colour in each trial. The target letter was to be regarded as a distracter if it occurred on non-target colour squares. We randomly assigned red, green, yellow or blue to be painted in each square of the 4×4 matrix. Initially, each colour display was shown for 267 ms and then to-be-searched letters and numbers appeared superimposed on the coloured squares for an additional 267 ms. If observers could fully utilize the colour information, they would only have to search one-quarter of the squares for the target, greatly simplifying the detection task. Whenever an observer's performance reached 95% correct, the display was speeded up and the duration of each frame was decreased. One-hour sessions were conducted every day until the performance of observers failed to improve in three consecutive runs.

Third-order motion paradigm

Observers reported the direction of motion in the third-order motion paradigm (Fig. 1b). Each trial contained five frames, with a 90° phase shift between consecutive frames. Even frames contained thin stripes of high-contrast noise alternating with thick stripes of low-contrast noise, whereas odd frames contained horizontal isoluminant red and green sine wave gratings (Fig. 1c). Red-green gratings were generated by modulating along the long-medium wavelength cardinal axis in Derrington-Krauskopf-Lennie colour space²³, and were calibrated carefully to eliminate the residual luminance. The first- and second-order motion systems are driven by stimuli (luminance and contrast/texture) that are very easy for our visual system to detect. Without careful calibration, the motion perception may be attributed to inputs from other stimulus components that are not in our interest to investigate, so it is important to ensure there is no contamination from those systems^{24,25}. However, even if some first- or second-order motion contamination remained, it would only make it more difficult to measure an effect of attention on motion, because contamination cannot produce a spurious attention effect.

In both even and odd frames, the mean luminance is constant and equal to the grey background. When the five frames are displayed successively, the two colours produce opposite, thus ambiguous, motion perception. The final perceived moving direction depends on the relative strength (salience) of the two colours, as there are no other cues (for example, luminance or contrast) known to contribute to motion perception. The salience of the two colours is jointly decided by their saturation and the attentional manipulation. If the colours are equally salient, ambiguous motion is perceived.

Procedures

A baseline psychometric function using the third-order motion paradigm was constructed before the search task. The percentage of trials in which observers reported motion in the red direction was plotted as a function of red-green saturation ratios (Fig. 2a, b, black line). After the search task was finished, another measurement was made with the same third-order motion paradigm and a new psychometric function was derived (coloured line in Fig. 2a, b). The lateral shift in the psychometric function indicated an increase in the salience of target colour. To measure the lifetime of the sensitization effect, we had four observers repeat the motion paradigm at regular intervals. A new psychometric function was constructed at each visit and compared with the baseline. We take the sum of differences from all points between the new and baseline psychometric functions as an index of the size of the sensitization effect. Positive values indicate increased sensitization to red, whereas negative values indicate sensitization to green. This index is shown as a function of time (Fig. 2c, d), and the data show that there is long-term sensitization for all our observers. This sensitization persists for weeks.

Received 13 September 2003; accepted 26 February 2004; doi:10.1038/nature02443.

- Blaser, E., Sperling, G. & Lu, Z.-L. Measuring the amplification of attention. *Proc. Natl Acad. Sci. USA* **96**, 11681-11686 (1999).
- Lu, Z.-L. & Sperling, G. The functional architecture of human visual motion perception. *Vision Res.* **35**, 2697-2722 (1995).

3. McCollough, C. Color adaptation of edge-detectors in the human visual system. *Science* **149**, 1115–1116 (1965).
4. Holland, H. C. *The Spiral After-Effect* (Pergamon, Oxford, 1965).
5. Blakemore, C. & Campbell, F. W. Adaptation to spatial stimuli. *J. Physiol. (Lond.)* **1**, 11–13 (1969).
6. Gilinsky, A. S. Orientation-specific effects of patterns of adapting light on visual acuity. *J. Opt. Soc. Am.* **58**, 13–18 (1968).
7. Wolfe, J. M. & O'Connell, K. M. Fatigue and structural change: Two consequences of visual pattern adaptation. *Investigative Ophthalmol. Vis. Sci.* **27**, 538–543 (1986).
8. Jones, P. D. & Holdings, D. H. Extremely long-term persistence of the McCollough effect. *J. Exp. Psychol.* **1**, 323–327 (1975).
9. Masland, R. H. Visual motion perception: Experimental modification. *Science* **165**, 819–821 (1969).
10. Blakemore, C., Nachmias, J. & Sutton, P. The perceived spatial frequency shift: Evidence for frequency-selective neurons in the human brain. *J. Physiol. (Lond.)* **210**, 727–750 (1970).
11. Kalfin, K. & Locke, S. Evaluation of long term visual motion after-image following monocular stimulation. *Vision Res.* **12**, 359–361 (1972).
12. Heggelund, P. & Krelling, S. Long-term retention of the Gilinsky-Effect. *Vision Res.* **16**, 1015–1017 (1976).
13. Favreau, O. E. Persistence of simple and contingent motion aftereffects. *Percept. Psychophys.* **26**, 187–194 (1979).
14. Neitz, J., Carroll, J., Yamauchi, Y., Neitz, M. & Williams, D. R. Color perception is mediated by a plastic neural mechanism that is adjustable in adults. *Neuron* **35**, 783–792 (2002).
15. Karni, A. & Sagi, D. The time course of learning a visual skill. *Nature* **365**, 250–252 (1993).
16. Fahle, M., Edelman, S. & Poggio, T. Fast perceptual learning in hyperacuity. *Vision Res.* **35**, 3003–3013 (1995).
17. Ramachandran, V. S. & Braddick, O. Orientation-specific learning in stereopsis. *Perception* **2**, 371–376 (1973).
18. McKee, S. P. & Westheimer, G. Improvement in vernier acuity with practice. *Percept. Psychophys.* **24**, 258–262 (1978).
19. Fiorentini, A. & Berardi, N. Perceptual learning specific for orientation and spatial frequency. *Nature* **287**, 43–44 (1980).
20. Ball, K. & Sekuler, R. A specific and enduring improvement in visual motion discrimination. *Science* **218**, 697–698 (1982).
21. Poggio, T., Fahle, M. & Edelman, S. Fast perceptual learning in visual hyperacuity. *Science* **256**, 1018–1021 (1992).
22. Karni, A. & Sagi, D. Where practice makes perfect in texture discrimination: evidence for primary visual cortex plasticity. *Proc. Natl Acad. Sci. USA* **88**, 4966–4970 (1991).
23. Derrington, A. M., Krauskopf, J. & Lennie, P. Chromatic mechanisms in lateral geniculate nucleus of Macaque. *J. Physiol.* **357**, 241–265 (1984).
24. Anstis, S. & Cavanagh, P. in *Color Vision* (eds Mollon, J. D. & Sharpe, E. T.) 155–166 (Academic, New York, 1983).
25. Lu, Z.-L. & Sperling, G. Sensitive calibration and measurement procedures based on the amplification principle in motion perception. *Vision Res.* **41**, 2355–2374 (2001).

Acknowledgements This research was supported by AFOSR, Life Science Directorate, Visual Information Processing Program.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to C.T. (CH_Tseng@alumni.uci.edu) or G.S. (Sperling@uci.edu).

IkB kinase- α acts in the epidermis to control skeletal and craniofacial morphogenesis

Alok K. Sil¹, Shin Maeda¹, Yuji Sano¹, Dennis R. Roop² & Michael Karin¹

¹Laboratory of Gene Regulation and Signal Transduction, Department of Pharmacology, School of Medicine, University of California, San Diego, 9500 Gilman Drive, La Jolla, California 92093-0636, USA

²Departments of Molecular and Cellular Biology and Dermatology, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA

IKB kinase- α (IKK- α)¹ exhibits protein-kinase-dependent and -independent functions. Its kinase activity is required for lymphoid organogenesis² and mammary gland development³, whereas a kinase-independent activity is required for epidermal keratinocyte differentiation⁴. In addition to failed epidermal differentiation, IKK- α -deficient mice exhibit abnormal skeletal and craniofacial morphogenesis^{4–6}. As similar defects are not exhibited by mice that experience systemic inhibition of NF- κ B⁷,

we postulated that the morphogenetic defects in IKK- α -deficient mice are not caused by reduced NF- κ B activity but instead are due to failed epidermal differentiation that disrupts proper epidermal–mesodermal interactions. We tested this hypothesis by introducing an epidermal-specific *Ikka* (also known as *Chuk*) transgene into IKK- α -deficient mice. Mice lacking IKK- α in all cell types including bone and cartilage, but not in basal epidermal keratinocytes, exhibit normal epidermal differentiation and skeletal morphology. Thus, epidermal differentiation is required for proper morphogenesis of mesodermally derived skeletal elements. One way by which IKK- α controls skeletal and craniofacial morphogenesis is by repressing expression of fibroblast growth factor (FGF) family members, such as FGF8, whose expression is specifically elevated in the limb bud ectoderm of IKK- α -deficient mice.

IKK- α -deficient (*Ikka*^{−/−}) mice, which die shortly after birth, exhibit marked morphological abnormalities that include taut and shiny skin, rudimentary limbs, absent or severely truncated tail, and a short and rounded head due to shorter jaw and nasal bones^{5,6,8}. Previous work revealed that IKK- α acts in a cell-autonomous manner to control terminal differentiation of epidermal keratinocytes⁴. This action of IKK- α does not depend on its protein kinase activity⁴, and *Ikka*^{AA} mice expressing an inactivatable form of IKK- α exhibit normal morphogenesis³. As epidermal–mesodermal interactions have a major role in vertebrate morphogenesis⁹, and as mice in which NF- κ B activation is inhibited do not exhibit skeletal abnormalities⁷, we examined whether failed keratinocyte differentiation underlies the morphogenetic defects exhibited by *Ikka*^{−/−} mice. This hypothesis was tested by introducing a human *Ikka* transgene driven by the basal keratinocyte-specific promoter of the cytokeratin 14 (*CK14*) gene¹⁰ (Fig. 1a) into *Ikka*^{−/−} mice. Several transgenic lines exhibiting germline transmission and efficient expression of the *CK14–Ikka* transgene were established and two of them were crossed with *Ikka*^{+/-} mice to generate *Ikka*^{−/−} *CK14–Ikka* progeny. Immunoblot analysis revealed efficient expression of IKK- α in the skin of these mice, but not in other tissues, including bone (Fig. 1b). Within the skin, expression of human IKK- α encoded by the *CK14–Ikka* transgene was restricted to the basal layer of the epidermis and hair follicles (Fig. 1c), consistent with the known specificity of the *CK14* promoter¹⁰. Immunoblot analysis of cultured keratinocytes revealed that the transgenic human IKK- α was expressed at levels similar to those of endogenous mouse IKK- α in wild-type keratinocytes (Supplementary Fig. 1).

Notably, the *CK14–Ikka* transgene completely rescued most of the morphological abnormalities exhibited by non-complemented *Ikka*^{−/−} mice (Fig. 1d). Furthermore, *Ikka*^{−/−} *CK14–Ikka* mice were covered by normal, wrinkled and loose skin, exhibited well-developed limbs and tail of normal length, and their head had a fully normal appearance, entirely different from that of non-complemented *Ikka*^{−/−} mice. Whereas *Ikka*^{−/−} mice died within the first hour after birth, *Ikka*^{−/−} *CK14–Ikka* mice survived for 1 day and during the first 12 h were as active as age-matched controls. Nonetheless, all *Ikka*^{−/−} *CK14–Ikka* mice died by the second day after birth, and visual examination revealed no milk in their stomachs (Fig. 1d). The cause for the suckling failure seems to be the result of a fused oesophagus, which is identical in appearance to that of non-complemented *Ikka*^{−/−} mice (Supplementary Fig. 2a). As with the epidermis, the upper part of the oesophagus contains a stratified epithelium¹¹, which fails to differentiate in *Ikka*^{−/−} mice⁸. Indeed, the *CK14–Ikka* transgene was not expressed in the oesophagus (Supplementary Fig. 2b). As CK14 is normally expressed in the oesophagus^{10,12}, the failure of transgene expression in this tissue may be due to the absence of essential regulatory elements. Histological, microscopic and immunohistochemical analyses revealed that the epidermis of *Ikka*^{−/−} *CK14–Ikka* mice was fully stratified, differentiated and indistinguishable from that of wild-type mice (Fig. 2a). In addition, both the epidermis and cultured epidermal keratinocytes

of the complemented mice expressed normal levels of filaggrin, a late-stage terminal differentiation marker (Fig. 2a, b). As the *Ikka*^{AA} mutant allele, which also allows normal epidermal differentiation³, encodes an IKK- α variant that exhibits basal kinase activity¹³, we repeated the rescue experiments using a *CK14-Ikka(KM)* transgene that codes for a catalytically inactive form of IKK- α ¹⁴. Again, complete rescue of epidermal differentiation was observed in *Ikka*^{-/-} mice that expressed this transgene (Supplementary Fig. 3).

Expression of the *CK14-Ikka* transgene also resulted in nearly complete reversal of all skeletal and craniofacial defects (Fig. 2c, d; see also Supplementary Fig. 4). The overall morphology of the skeleton and skull was indistinguishable between wild-type and *Ikka*^{-/-} *CK14-Ikka* mice (Fig. 2c). The skull, rib cage, limb bones and tail of the transgene-rescued mice were of normal size and morphology. The syndactyly and phalanx truncations of both fore- and hindlimbs exhibited by *Ikka*^{-/-} mice were completely prevented (Fig. 2d; see also Supplementary Fig. 4a). Close examination of the skull failed to reveal any obvious differences between wild-type and *Ikka*^{-/-} *CK14-Ikka* or *Ikka*^{-/-} *CK14-Ikka(KM)* mice (Supplementary Fig. 5). In addition, the transgenic rescue resulted in complete fusion of the sternum and xiphoid process as opposed to the split sternum and bifurcated xiphoid of *Ikka*^{-/-} mice (Supplementary Fig. 4b). The only minor skeletal defect present in the rescued mice was fusion of the last two pairs of ribs (Supplementary Fig. 4b). In addition to normal skeletal and craniofacial morphology, it should be noted that the pattern of ossification was essentially identical between wild-type and *Ikka*^{-/-} *CK14-Ikka* mice (Fig. 2c, d; see also Supplementary Fig. 4).

These results indicate that IKK- α -induced differentiation of epidermal keratinocytes is required for normal morphogenesis of

several mesodermally derived skeletal elements, including skull, sternum, digits and tail. This raises the question of how the epidermis controls the morphogenesis of the underlying mesoderm. Early studies of limb and ectodermal appendage development revealed that the mesoderm controls epidermal morphogenesis⁹. More recent experiments, however, have outlined reciprocal interactions between ectodermal- and mesodermal-derived signalling and patterning molecules¹⁵. Many of the skeletal and craniofacial defects displayed by *Ikka*^{-/-} mice resemble those caused by

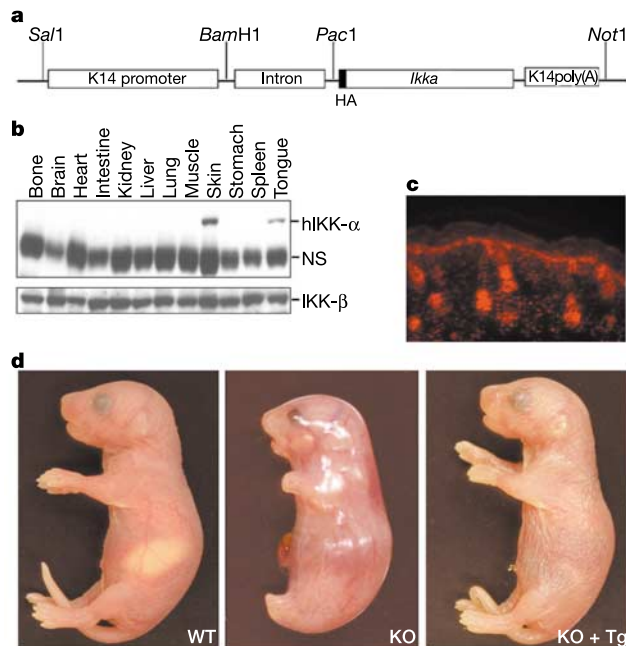


Figure 1 The *CK14-Ikka* transgene restores normal physical appearance to *Ikka*^{-/-} mice. **a**, Schematic drawing of the *CK14-Ikka* transgene. **b**, Immunoblot analysis of transgenic human (h) IKK- α expression in different tissues of *Ikka*^{-/-} *CK14-Ikka* mice using an antibody that recognizes both human and mouse IKK- α . Expression of transgenic human IKK- α was detected only in the skin and tongue. No mouse IKK- α was detected. IKK- β and a nonspecific (NS) band were used as loading controls. **c**, Immunohistochemical analysis of transgenic human IKK- α expression in mouse skin. Skin sections from *Ikka*^{-/-} *CK14-Ikka* (KO + Tg) mice were immunostained with anti-IKK- α antibody. Expression of transgenic human IKK- α was detected in the basal layer of the epidermis and in hair follicles. **d**, Phenotypes of newborn wild-type (WT), *Ikka*^{-/-} (KO) and *Ikka*^{-/-} *CK14-Ikka* mice.

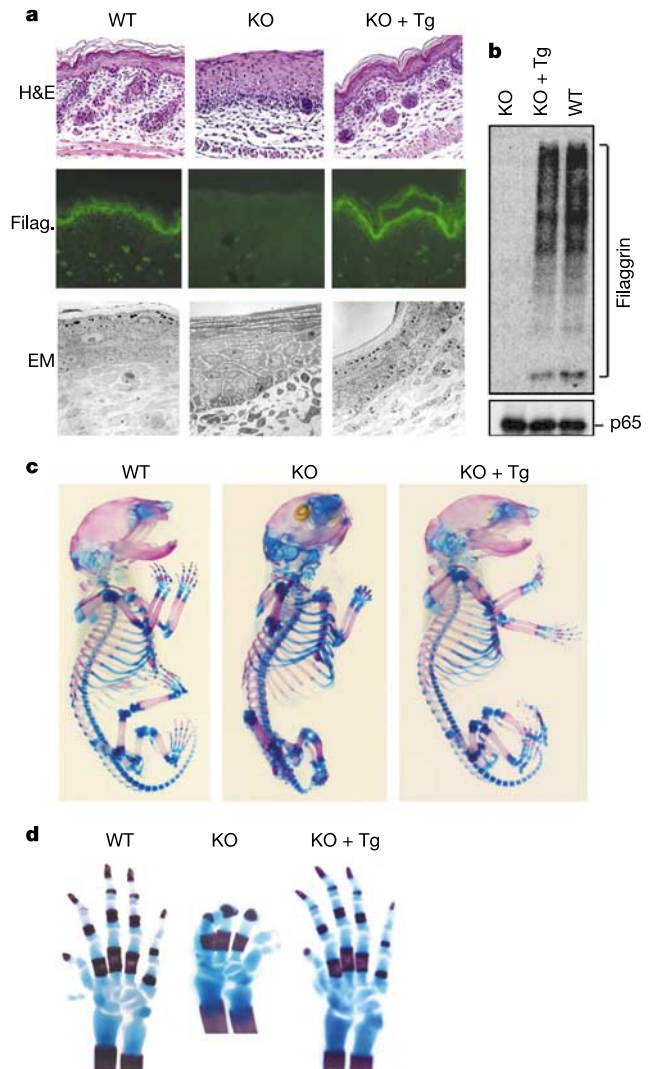


Figure 2 The *CK14-Ikka* transgene rescues skin as well as craniofacial and skeletal defects of *Ikka*^{-/-} mice. **a**, Abnormal epidermal development of *Ikka*^{-/-} mice is rescued by the *CK14-Ikka* transgene. Dorsal skin from newborn wild-type, *Ikka*^{-/-} and *Ikka*^{-/-} *CK14-Ikka* mice was sectioned and stained with haematoxylin and eosin (H&E) and anti-filaggrin antibody. Mid-thoracic skin from newborn mice of the different genotypes was fixed and examined by electron microscopy (EM; magnification $\times 1,000$) after staining with lead citrate. All three analyses reveal that the *CK14-Ikka* transgene completely rescues epidermal differentiation, leading to a fully stratified, terminally differentiated epithelium. **b**, Normal differentiation of cultured *Ikka*^{-/-} *CK14-Ikka* keratinocytes. Primary cultures of keratinocytes from wild-type, *Ikka*^{-/-} and *Ikka*^{-/-} *CK14-Ikka* mice were induced to differentiate by reaching high density. Protein lysates were prepared and examined for expression of filaggrin by immunoblotting. RelA/p65 was used as a loading control. **c**, Normal skeleton of a *Ikka*^{-/-} *CK14-Ikka* mouse. Skeletons of newborn wild-type, *Ikka*^{-/-} and *Ikka*^{-/-} *CK14-Ikka* mice were double-stained with alizarin red and alcian blue. Bone is stained red and cartilage is blue. Magnified forelimbs are shown in **d**.

deficiencies in bone morphogenic protein (BMP) signalling or by activating mutations in FGF receptors (FGFR). For instance, split sternum and bifurcated xiphoid are exhibited by *Bmp5*^{-/-} *Bmp6*^{-/-} doubly deficient mice¹⁶, whereas syndactyly and phalanx truncations are exhibited by mice lacking BMP receptor 1b (BMPRIb)¹⁷. Tail truncations are seen in chickens expressing dominant negative BMPRI (ref. 18) and in mice lacking BMP4 (ref. 19). Craniofacial and skeletal abnormalities including mid-face underdevelopment—similar to those exhibited by *Ikka*^{-/-} mice—and digit truncations are exhibited by human patients with activating mutations in FGFRs²⁰. Thus, a possible mechanism accounting for abnormal morphogenesis in *Ikka*^{-/-} mice may include decreased production of BMPs or increased production of FGFs, which antagonize BMP signalling²¹.

This hypothesis was examined by analysing RNA samples of fore- and hindlimbs from embryonic day 13 (E13) wild-type and *Ikka*^{-/-} fetuses by real-time polymerase chain reaction (PCR). Although no significant changes were observed in expression of BMP1, -2, -4, -5, -6 and -7, a modest increase in expression of the BMP antagonist gremlin was observed in *Ikka*^{-/-} limbs, whereas expression of another BMP inhibitor, follistatin, remained unchanged (Fig. 3a). More striking were the alterations in expression of FGF family members. Whereas expression of FGF2, -4 and -7 was unchanged, a modest elevation in expression of FGF6 and -10 was observed, and expression of FGF3, -5, -15 and -17 was strongly elevated, with the most marked elevation in *Ikka*^{-/-} mice observed for FGF8 messenger RNA (Fig. 3a). Introduction of the *CK14-Ikka* transgene reversed all of these changes (Fig. 3b). Upregulation of FGF8 mRNA in *Ikka*^{-/-} embryos was confirmed by *in situ* hybridization, which

revealed accumulation of FGF8 mRNA at the outer edges of limb buds and head of E12.5 to E13 fetuses (Fig. 3c). Examination of limb bud sections revealed FGF8 mRNA accumulation in the ectoderm. Using cultured keratinocytes from wild-type, *Ikka*^{-/-} and *Ikka*^{-/-} *CK14-Ikka* neonate mice, we detected elevated expression of FGF2, -3, -7, -10 and -18 in *Ikka*^{-/-} keratinocytes, with normal expression levels after introduction of the *CK14-Ikka* transgene (Fig. 3d). Consistent with FGF8 overexpression, we found elevated expression of two FGF-inducible genes, *Mkp3* (ref. 22) and *Spry1* (ref. 23), in E13 *Ikka*^{-/-} fetuses (Fig. 3e). Overexpression of both genes was reversed by the *CK14-Ikka* transgene. FGFR activation should stimulate mitogen-activated protein kinases such as ERK. Indeed, we found elevated ERK phosphorylation in *Ikka*^{-/-} limb buds (Fig. 3f). As non-differentiated wild-type keratinocytes expressed higher levels of FGF family members than differentiated keratinocytes (Supplementary Fig. 6), it is possible that the effect of IKK-α on FGF gene expression is indirect.

To determine whether increased FGFR activity is responsible for the defects in limb development, we established an explant culture system²⁴. Explanted wild-type E12 limb buds exhibited clear digitation (Fig. 4). By contrast, *Ikka*^{-/-} limb buds did not display digitation at any time point, although their survival *ex vivo* was similar to wild-type counterparts (Fig. 4). However, incubation of *Ikka*^{-/-} limb buds with the FGFR-specific tyrosine kinase inhibitor SU5402 (ref. 25) resulted in appearance of interdigital separations. The inhibitor had no effect on digitation in wild-type limb buds.

As mentioned above, IKK-α may regulate FGF mRNA levels through its ability to induce keratinocyte differentiation. To gain further insight into this process, we examined the subcellular

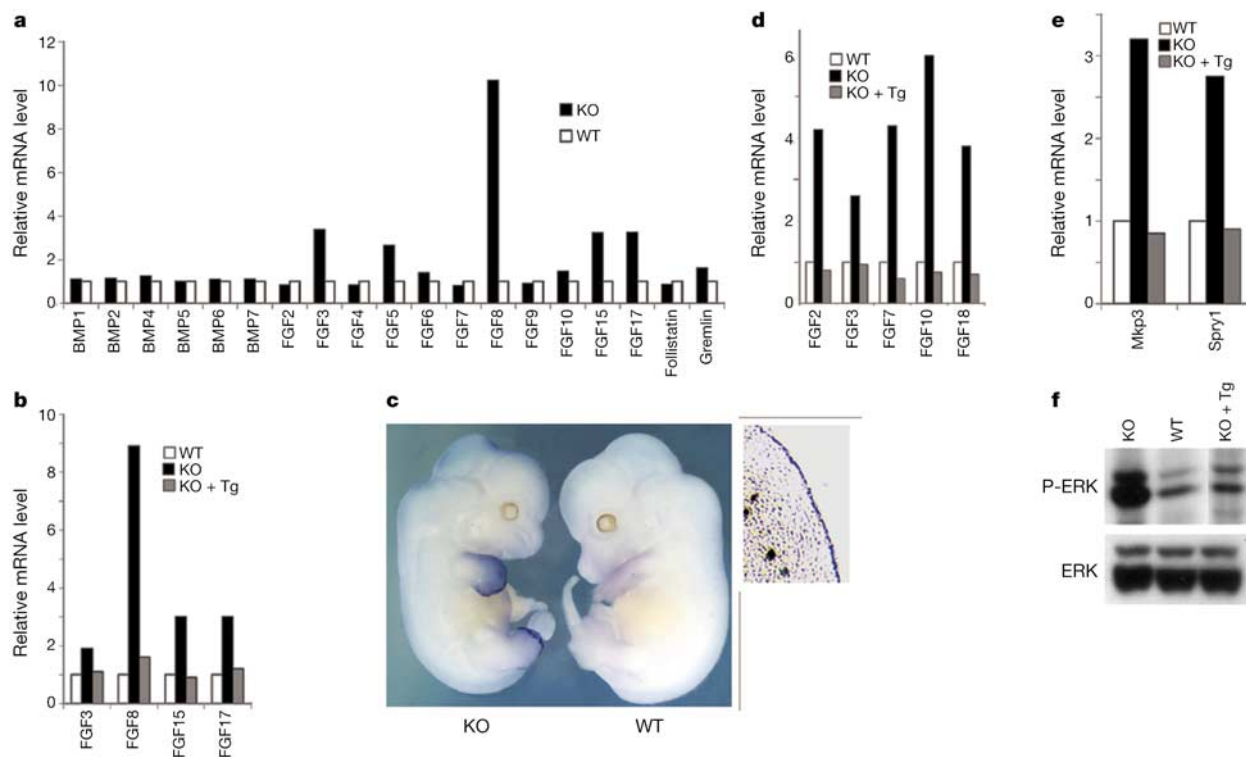


Figure 3 Elevated FGF mRNA expression in *Ikka*^{-/-} fetal limbs. **a**, Analysis of gene expression. Total RNAs from limbs of wild-type and *Ikka*^{-/-} E13 embryos were analysed by real-time PCR for expression of mRNAs coding for FGFs, BMPs and BMP antagonists. Results are averages of two separate experiments, normalized to the level of cyclophilin mRNA. **b**, The *CK14-Ikka* transgene suppresses FGF mRNA overexpression. Total RNA was isolated from limbs of E13 wild-type, *Ikka*^{-/-} and *Ikka*^{-/-} *CK14-Ikka* mice and analysed as above. **c**, Analysis of FGF mRNA expression by whole-mount *in situ* hybridization of E12.5–E13 embryos. The side panel shows a section of the *Ikka*^{-/-} limb

bud demonstrating elevated FGF8 mRNA at the distal ectoderm. **d**, Elevated FGF mRNA expression in *Ikka*^{-/-} cultured keratinocytes. Total RNA was isolated from 1-week-old primary keratinocyte cultures derived from newborn wild-type, *Ikka*^{-/-} and *Ikka*^{-/-} *CK14-Ikka* mice. **e**, Elevated *Mkp3* and *Spry1* mRNAs in *Ikka*^{-/-} limbs. RNA analysis was carried out as in **a**. **f**, Elevated mitogen-activated protein kinase activity in *Ikka*^{-/-} limb buds. Lysates of E13 limbs were gel-separated and immunoblotted with antibodies that recognize total ERK1/2 and their activated form (P-ERK).

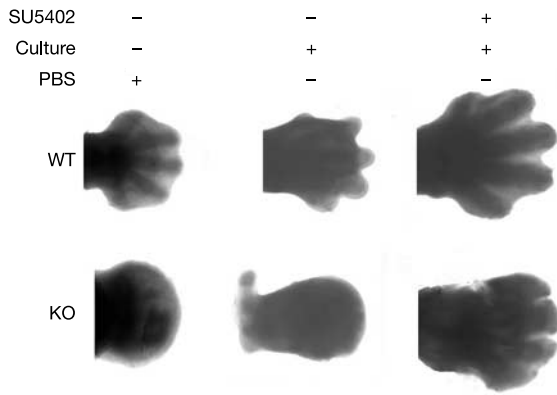


Figure 4 FGFR inhibitor rescues digitation in *Ikka*^{-/-} limb buds. Limb buds were isolated from E12 embryos and cultured in ascorbic-acid-supplemented BGJB medium. A set of limb buds was kept in PBS for negative control. After 20 h, the FGFR inhibitor SU5402 was added at 1.5 mg ml⁻¹ and limbs were allowed to grow for another 20 h. Wild-type limb buds exhibit digitation in the absence or presence of SU5402, whereas *Ikka*^{-/-} limb buds exhibit digitation only after incubation with the inhibitor. This experiment was performed several times using limb buds from E12, E13.5 and E14.5 embryos with identical results.

distribution of IKK- α in non-differentiated and differentiated keratinocytes, and found that induction of keratinocyte differentiation was associated with higher levels of nuclear IKK- α (Fig. 5a). Sequence analysis revealed the presence of a nuclear localization sequence (NLS) within the kinase domain of IKK- α (Fig. 5b). Such a sequence is not present in IKK- β , which was never found in the nucleus (data not shown). Inactivation of the NLS by site-directed mutagenesis (Fig. 5b) prevented the ability of IKK- α to enter the nucleus but did not interfere with its kinase activity or binding to

IKK- γ (Fig. 5c). Notably, the NLS mutant was incapable of inducing differentiation of *Ikka*^{-/-} keratinocytes (Fig. 5d). These results indicate that IKK- α acts in the nucleus to induce keratinocyte differentiation. This function does not depend on the kinase activity of IKK- α .

Our experiments suggest that the major site of IKK- α action is the basal keratinocyte of the epidermis, in which it is required for cell cycle exit and terminal differentiation⁴. This activity of IKK- α is independent of its protein kinase activity and is probably exerted within the nucleus. Thus, despite a recent suggestion—based on analysis of immortalized (and hence abnormal) fibroblast cell lines derived from *Ikka*^{-/-} mice—that IKK- α may function as a histone H3 kinase²⁶, this function is not required for induction of keratinocyte differentiation. Furthermore, the absence of liver apoptosis in both *Ikka*^{-/-} (refs 5, 6, 8) or *Ikka*^{-/-} CK14-*Ikka* mice strongly suggests that IKK- α does not have a critical role in TNF- α -induced NF- κ B-dependent gene transcription, at least in hepatocytes. Nevertheless, through its ability to induce keratinocyte differentiation, IKK- α exerts a strong effect on morphogenesis of several mesodermally derived skeletal structures, such as the skull, sternum, digits and tail. One mechanism by which keratinocyte-specific synthesis of IKK- α affects craniofacial and skeletal morphogenesis is through regulation of FGF family members. FGFs and their receptors exert marked effects on craniofacial and skeletal morphogenesis^{20,21}. Several FGFs are expressed at different time points in the epidermis and can antagonize the function of BMPs through a sophisticated relay system^{21,27}, and they also inhibit apoptosis through activation of anti-apoptotic signalling molecules²². Thus, overproduction of FGFs, especially FGF8, by non-differentiated IKK- α -deficient keratinocytes can account for many of the skeletal and craniofacial defects in *Ikka*^{-/-} mice, and inhibition of FGFR kinase activity results in appearance of interdigital spaces in *Ikka*^{-/-} limb buds. By providing a clear and genetically based demonstration of the ability of the epidermis to control the morphogenesis of mesodermally

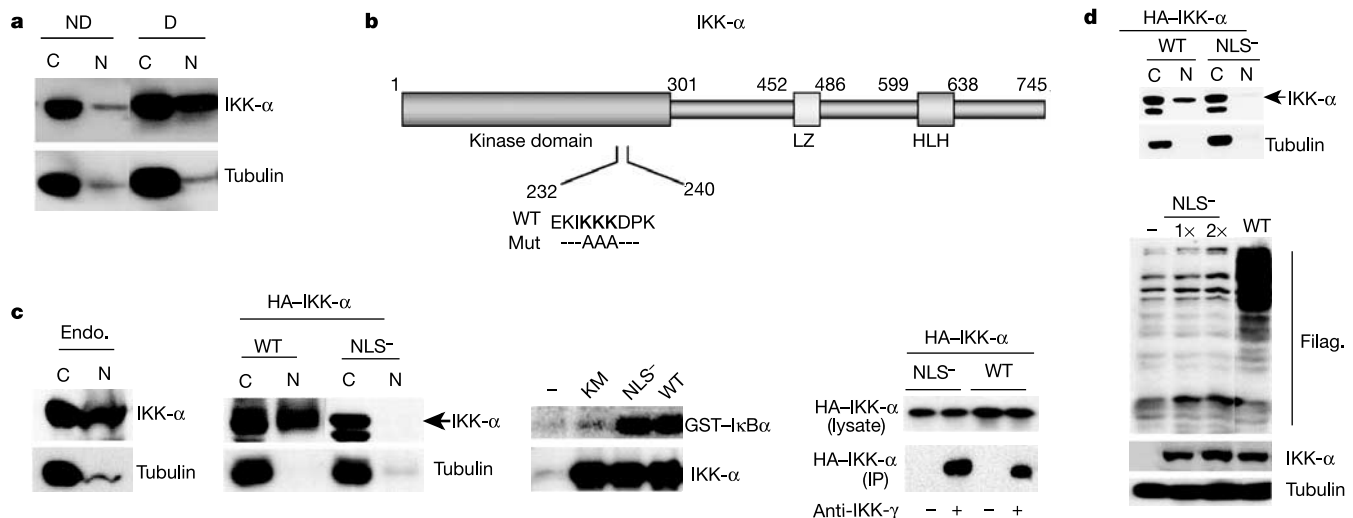


Figure 5 IKK- α acts within the nucleus to induce keratinocyte differentiation. **a**, IKK- α translocates to the nucleus of differentiating keratinocytes. Nuclear (N)–cytosolic (C) fractionation was performed on non-differentiated (ND) and differentiated (D) keratinocytes. Distribution of IKK- α and α -tubulin was examined by immunoblotting. **b**, Schematic diagram of IKK- α and its putative NLS. Residues altered by site-directed mutagenesis are indicated. **c**, Characterization of the IKK- α NLS mutant. First panel: endogenous IKK- α is detected in both cytosolic and nuclear fractions of HeLa cells. Second panel: the IKK- α NLS mutant does not enter the nucleus. HeLa cells were transiently transfected with wild-type and NLS mutant IKK- α haemagglutinin (HA)-tagged expression vectors. The subcellular distribution of the HA-tagged proteins was examined as above. Third panel: the IKK- α NLS mutant has normal kinase activity. Wild-type and NLS mutant IKK- α were expressed in HEK293 cells and their kinase activity was examined

using an immune complex kinase assay with glutathione *S*-transferase (GST)–I κ B α as a substrate. Fourth panel: NLS mutant IKK- α binds to IKK- γ . HEK293 cells were transiently transfected with wild-type and mutant IKK- α expression vectors. Cell lysates were immunoprecipitated with anti-IKK- γ antibody and the co-precipitation of wild-type and mutant NLS IKK- α was analysed by immunoblotting with anti-HA antibody. **d**, Nuclear entry of IKK- α is required for keratinocyte differentiation. Primary cultures of *Ikka*^{-/-} keratinocytes were infected with adenoviruses expressing either wild-type or NLS mutant IKK- α . Top panel: nuclear and cytoplasmic extracts were examined for presence of IKK- α and α -tubulin. Bottom panel: cell lysates were prepared 48 h after infection and differentiation was examined by immunoblotting for filaggrin. Expression of wild-type and NLS mutant IKK- α and the level of α -tubulin (loading control) were also monitored by immunoblotting.

derived craniofacial and skeletal elements, our work complements early grafting experiments that demonstrated the influence of the mesoderm on epidermal differentiation⁹. □

Methods

Generation of *CK14-Ikka* transgenic mice

A *HindIII*-*NotI* fragment encoding human *Ikka* was subcloned into the *SnaBI* site of pBS-R3, which contains the human *CK14* promoter. The *SalI*-*NotI* fragment containing a *CK14-Ikka* poly(A) cassette was isolated, purified and microinjected into male pro-nuclei of fertilized C57BL/6 oocytes. The *CK14-Ikka(KM)* construct was similarly generated except that the expression cassette was injected into oocytes isolated from an *Ikka*^{+/-} intercross. Transgenic founders were identified by PCR analysis and confirmed by immunoblotting. Subsequently, two different lines of transgenic founders were crossed with *Ikka*^{+/-} mice. *Ikka*^{+/-} *CK14-Ikka* mice were crossed to generate rescued knockout mice. *Ikka*^{-/-} *CK14-Ikka(KM)* mice were identified by genotyping of tissues derived from neonate mice.

Immunoblotting and immunohistochemistry

Protein lysates were prepared from different tissues of *Ikka*^{-/-} *CK14-Ikka* newborn mice and immunoblotted with an anti-IKK- α antibody (Imgenex) and anti-IKK- β antibody (USB) (Fig. 1b). Protein extracts prepared from primary keratinocyte cultures (5–7 days old) were analysed by immunoblotting with an anti-IKK- α antibody (Imgenex) that recognizes both human and mouse IKK- α (Supplementary Fig. 1) or with an anti-flaggrin antibody (Babco) (Fig. 2b).

Back skin and oesophagus from newborn mice were fixed in formalin and Bouin's solution, respectively. Fixed tissues were paraffin-embedded, serially sectioned at 5 μ m and stained with haematoxylin and eosin. Immunohistochemical staining for flaggrin was performed on paraffin sections with anti-flaggrin antibody (Babco). Immunohistochemical staining for IKK- α was performed on 4% paraformaldehyde-fixed cryosections using anti-IKK- α antibody (Cell Signalling) as per the manufacturer's recommendation. Bone and cartilage staining of newborn mice was performed using alcian blue and alizarin red as described previously²⁸.

In situ hybridization

Whole-mount *in situ* hybridization was performed on 4% paraformaldehyde-fixed E12.5 embryos as described²⁹. A digoxigenin-labelled anti-sense mouse FGF8 probe, corresponding to the coding region, was used.

In vitro culture of mouse limb buds

Isolated embryonic limb buds were cultured as described²⁴ using serum-free, defined BGJB medium supplemented with 0.2 mg ml⁻¹ ascorbic acid and penicillin–streptomycin.

Real-time PCR analysis

Real-time PCR was performed as described^{3,30}. Total cellular RNA used for this analysis was isolated from limbs of E13 embryos. Cyclophilin mRNA was used to normalize the amount and quality of the RNA.

Received 29 December 2003; accepted 12 February 2004; doi:10.1038/nature02421.

- DiDonato, J. A., Hayakawa, M., Rothwarf, D. M., Zandi, E. & Karin, M. A cytokine-responsive I κ B kinase that activates the transcription factor NF- κ B. *Nature* **388**, 548–554 (1997).
- Senfleben, U. *et al.* Activation by IKK α of a second, evolutionary conserved, NF- κ B signaling pathway. *Science* **293**, 1495–1499 (2001).
- Cao, Y. *et al.* IKK α provides an essential link between RANK signaling and cyclin D1 expression during mammary gland development. *Cell* **107**, 763–775 (2001).
- Hu, Y. *et al.* IKK α controls formation of the epidermis independently of NF- κ B via a differentiation inducing factor. *Nature* **410**, 710–714 (2001).
- Takeda, K. *et al.* Limb and skin abnormalities in mice lacking IKK α . *Science* **284**, 313–316 (1999).
- Li, Q. *et al.* IKK1-deficient mice exhibit abnormal development of skin and skeleton. *Genes Dev.* **13**, 1322–1328 (1999).
- Schmidt-Ullrich, R. *et al.* Requirement of NF- κ B/Rel for the development of hair follicles and other epidermal appendages. *Development* **128**, 3843–3853 (2001).
- Hu, Y. *et al.* Abnormal morphogenesis but intact IKK activation in mice lacking the IKK α subunit of the I κ B kinase. *Science* **284**, 316–320 (1999).
- Wessells, N. K. *Tissue Interactions and Development* (Benjamin-Cummings, Menlo Park, California, 1997).
- Vassar, R., Rosenberg, M., Ross, S., Tyner, A. & Fuchs, E. Tissue-specific and differentiation-specific expression of a human K14 keratin gene in transgenic mice. *Proc. Natl Acad. Sci. USA* **86**, 1563–1567 (1989).
- Torrey, T. W. *Morphogenesis of the Vertebrates* 275–308 (Wiley, New York, 1970).
- Takahashi, H., Shikata, N., Senzaki, H., Shintaku, M. & Tsubura, A. Immunohistochemical staining patterns of keratins in normal oesophageal epithelium and carcinoma of the oesophagus. *Histopathology* **26**, 45–50 (1995).
- Delhase, M., Hayakawa, M., Chen, Y. & Karin, M. Positive and negative regulation of I κ B kinase activity through IKK β subunit phosphorylation. *Science* **284**, 309–313 (1999).
- Zandi, E., Chen, Y. & Karin, M. Direct phosphorylation of I κ B by IKK α and IKK β : Discrimination between free and NF- κ B-bound substrate. *Science* **281**, 1360–1363 (1998).
- Dudley, A. T. & Tabin, C. J. Constructive antagonism in limb development. *Curr. Opin. Genet. Dev.* **10**, 387–392 (2000).
- Solloway, M. J. *et al.* Mice lacking Bmp6 function. *Dev. Genet.* **22**, 321–339 (1998).
- Yi, S. E., Daluiski, A., Pederson, R., Rosen, V. & Lyons, K. M. The type I BMP receptor BMPRII is required for chondrogenesis in the mouse limb. *Development* **127**, 621–630 (2000).
- Zou, H. & Niswander, L. Requirements for BMP signaling in interdigital apoptosis and scale formation. *Science* **272**, 738–741 (1996).

- Winnier, G., Blessing, M., Labosky, P. A. & Hogan, B. L. Bone morphogenetic protein-4 is required for mesoderm formation and patterning in the mouse. *Genes Dev.* **9**, 2105–2116 (1995).
- Webster, M. K. & Donoghue, D. J. FGFR activation in skeletal disorders: too much of a good thing. *Trends Genet.* **13**, 178–182 (1997).
- Martin, G. R. The roles of FGFs in the early development of vertebrate limbs. *Genes Dev.* **12**, 1571–1586 (1998).
- Kawakami, Y. *et al.* MKP3 mediates the cellular response to FGF8 signalling in the vertebrate limb. *Nature Cell Biol.* **5**, 513–519 (2003).
- Minowada, G. *et al.* Vertebrate Sprouty genes are induced by FGF signaling and can cause chondrodysplasia when overexpressed. *Development* **126**, 4465–4475 (1999).
- Zuniga, A., Haramis, A. P., McMahon, A. P. & Zeller, R. Signal relay by BMP antagonism controls the SHH/FGF4 feedback loop in vertebrate limb buds. *Nature* **401**, 598–602 (1999).
- Mohammadi, M. *et al.* Structures of the tyrosine kinase domain of fibroblast growth factor receptor in complex with inhibitors. *Science* **276**, 955–960 (1997).
- Yamamoto, Y. *et al.* A nucleosomal role of IKK α is critical for cytokine-induced activation of NF- κ B regulated genes. *Nature* **423**, 655–659 (2003).
- Welscher, P. *et al.* Progression of vertebrate limb development through SHH-mediated counteraction of GLI3. *Science* **298**, 827–830 (2002).
- Martin, J. F., Bradley, A. & Olson, E. N. The *paired-like* homeo box gene *MHox* is required for early events of skeletogenesis in multiple lineages. *Genes Dev.* **9**, 1237–1249 (1995).
- Wilkinson, D. G. *In Situ Hybridization* (Oxford Univ. Press, Oxford, 1993).
- Chen, L.-W. *et al.* The two faces of IKK and NF- κ B inhibition: Prevention of systemic inflammation but increased local injury following intestinal ischemia-reperfusion. *Nature Med.* **9**, 575–581 (2003).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank A. Ullrich for the FGF receptor inhibitor (SU5402), T. Kato, M. Delhase and S. Roy for technical assistance and discussions, and M. Ellisman and M. Mackey for the electron microscopy performed at the National Center for Microscopy and Imaging Research. Work was supported by grants from the National Institutes of Health, and Superfund Basic Research Program and CERES research awards.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.K. (karinoffice@ucsd.edu).

Haematopoietic stem cells do not transdifferentiate into cardiac myocytes in myocardial infarcts

Charles E. Murry¹, Mark H. Soonpaa², Hans Reinecke¹, Hidehiro Nakajima², Hisako O. Nakajima², Michael Rubart², Kishore B. S. Pasumarthi^{2*}, Jitka Ismail Virag¹, Stephen H. Bartelmez³, Veronica Poppa¹, Gillian Bradford², Joshua D. Dowell², David A. Williams^{2*} & Loren J. Field²

¹Department of Pathology, Box 357470, Room D-514 HSB, University of Washington, Seattle, Washington 98195, USA

²Wells Center for Pediatric Research, Indiana University, 1044 West Walnut Street, R4 Bldg, Room W376, Indianapolis 46202-5225, USA

³Department of Pathobiology, University of Washington, Seattle, Washington 98195, USA

* Present addresses: Department of Pharmacology Dalhousie University, Sir Charles Tupper Medical Bldg, Room 6-F1, 5850 College Street, Halifax B3H 1X5, Canada (K.B.S.P.); Division of Experimental Hematology, Cincinnati Children's Hospital Medical Center, University of Cincinnati College of Medicine, 3333 Burnet Avenue, Cincinnati, Ohio 45229-3039, USA (D.A.W.)

The mammalian heart has a very limited regenerative capacity and, hence, heals by scar formation¹. Recent reports suggest that haematopoietic stem cells can transdifferentiate into unexpected phenotypes such as skeletal muscle^{2,3}, hepatocytes⁴, epithelial cells⁵, neurons^{6,7}, endothelial cells⁸ and cardiomyocytes^{8,9}, in response to tissue injury or placement in a new environment. Furthermore, transplanted human hearts contain myocytes derived from extra-cardiac progenitor cells^{10–12}, which may have originated from bone marrow^{8,13–15}. Although most studies suggest that transdifferentiation is extremely rare under physiological conditions, extensive regeneration of myocardial infarcts

was reported recently after direct stem cell injection⁹, prompting several clinical trials^{16,17}. Here, we used both cardiomyocyte-restricted and ubiquitously expressed reporter transgenes to track the fate of haematopoietic stem cells after 145 transplants into normal and injured adult mouse hearts. No transdifferentiation into cardiomyocytes was detectable when using these genetic techniques to follow cell fate, and stem-cell-engrafted hearts showed no overt increase in cardiomyocytes compared to sham-engrafted hearts. These results indicate that haematopoietic stem cells do not readily acquire a cardiac phenotype, and raise a cautionary note for clinical studies of infarct repair.

A transgenic mouse line in which the cardiac-specific α -myosin heavy chain promoter drives expression of a nuclear-localized β -galactosidase reporter was used to monitor cardiomyogenic transdifferentiation events. These mice, designated as MHC-nLAC, have been used previously as donors for fetal cardiomyocyte transplantation¹⁸, where the robust nuclear β -galactosidase signal readily permitted detection of grafted cardiomyocytes in a wild-type heart after 5-bromo-4-chloro-3-indolyl- β -D-galactoside (X-gal) staining (Fig. 1a, b). Bone-marrow-derived haematopoietic stem cells (HSCs) were obtained from the MHC-nLAC mice by first isolating unfractionated marrow cells, and then removing cells expressing differentiated haematopoietic lineage cell-surface markers followed by fluorescence-activated cell sorting (FACS) to isolate cells expressing c-kit. The resulting population of primitive 'Lin⁻ c-kit⁺' cells, which are enriched in HSCs, was transplanted into the peri-infarct zone of adult congenic, non-transgenic recipients 5 h after coronary artery occlusion ($n = 42$). Our occlusion protocol typically resulted in infarct sizes of $38 \pm 5\%$ of the left ventricle¹⁹. The Lin⁻ c-kit⁺ stem cells were also transplanted into hearts injured by cauterization ($n = 26$), where the volume of injured myocardium is considerably less²⁰. Mice were killed (1–4

weeks after infarction; Table 1), and the hearts were fixed and vibratome-sectioned at 300 μ m from apex to base. The sections were then stained with X-gal and examined under a stereomicroscope. Despite the ability of this assay to detect a single transplanted fetal cardiomyocyte²¹, no blue nuclei were detected in any of the hearts that received stem cell transplants (Fig. 1c and Table 1). Furthermore, immunostaining of these hearts revealed no ectopic expression of sarcomeric myosin heavy chain in the infarcts (Fig. 1d). These observations suggest that the transplanted Lin⁻ c-kit⁺ HSCs had not transdifferentiated into cardiomyocytes.

In vitro co-culture experiments were performed to explore further the cardiogenic potential of the Lin⁻ c-kit⁺ cells prepared from MHC-nLAC mice. Hanging-drop cultures were used to generate chimaeric embryoid bodies derived in part from non-transgenic mouse embryonic stem (ES) cells and in part from Lin⁻ c-kit⁺ HSCs prepared from MHC-nLAC mice ($n = 840$ embryoid bodies). Chimaeric embryoid bodies were also generated by bulk mixing of the ES and HSCs ($n =$ approximately 400 embryoid bodies). A similar approach was used previously to monitor transdifferentiation of adult-derived neuronal stem cells²². After 3 days of suspension culture, the embryoid bodies were transferred to adherent surfaces and allowed to grow for an additional 10 days, at which time widespread spontaneous contractile activity was present. The cultures were then fixed and stained with X-gal. Out of the more than 1,000 attached chimaeric embryoid bodies screened, no blue nuclei were identified (Fig. 1e). For each experiment, regions of the dish with contractile activity were microdissected with a Pasteur pipette, and DNA was prepared and subjected to polymerase chain reaction (PCR) amplification (Fig. 1e). The presence of the MHC-nLAC reporter gene was readily detected, indicating that MHC-nLAC HSCs were present but failed to give rise to overt cardiac transdifferentiation events after place-

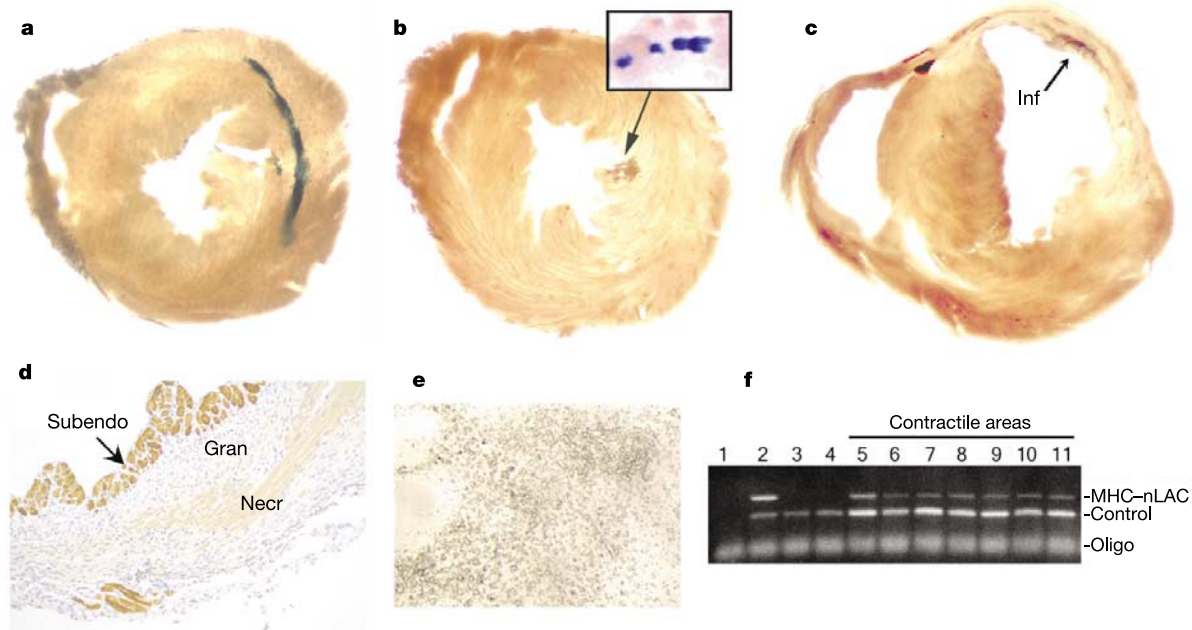


Figure 1 Failure of HSCs from MHC-nLAC mice to activate cardiac reporter genes or express endogenous myosin heavy chain. **a, b**, Positive control vibratome sections of hearts engrafted with fetal MHC-nLAC cardiomyocytes showing a large (**a**) and small (**b**) graft. Insert in **b** shows that a small cluster of donor cardiomyocytes is easily detectable despite fibrosis. **c**, X-gal-stained vibratome section from an infarcted heart receiving 100,000 MHC-nLAC Lin⁻ c-kit⁺ stem cells. No reporter gene activation is present. Inf, infarct. **d**, Histological section from an infarcted heart receiving 100,000 MHC-nLAC Lin⁻ c-kit⁺ HSCs and immunostained for sarcomeric myosin heavy chain. No myosin heavy chain is detected in the infarcted zone. Subendo, spared subendocardial

myocardium; Gran, granulation tissue; Necr, unresorbed necrotic myocardium.

e, Contractile region from a chimaeric embryoid body containing non-transgenic ES cells and MHC-nLAC Lin⁻ c-kit⁺ HSCs. No evidence for cardiomyogenic induction is apparent after X-gal staining. **f**, PCR from beating foci in chimaeric embryoid bodies demonstrating the presence of MHC-nLAC Lin⁻ c-kit⁺ cells. Lane 1, negative control without DNA; lane 2, positive control from a MHC-nLAC transgenic mouse; lane 3, negative control from a non-transgenic mouse; lane 4, negative control from an embryoid body with non-transgenic ES cells; lanes 5–11, PCR from microdissected contractile regions of chimaeric embryoid bodies containing MHC-nLAC Lin⁻ c-kit⁺ HSCs.

Table 1 Summary of intracardiac HSC transplantation data

Donor cell genotype	Haematopoietic stem cell used	Heart injury	Number of cells transplanted	Graft age at death (days)	Cardiomyogenic events per graft
MHC-nLAC	Lin ⁻ c-kit ⁺	MI	100,000	14–28	0/42
	Lin ⁻ c-kit ⁺	Cautery	100,000	7–26	0/26
	Lin ⁻ c-kit ⁺ Sca-1 ⁺	Cautery	31,000–75,000	7–36	0/11
	Lin ⁻ c-kit ⁺ Sca-1 ⁺	None	40,000–65,000	1–119	0/9
	Lin ⁻ c-kit ⁺ Sca-1 ⁺	Cautery	17,000–25,000	7–36	0/11
	Lin ⁻ c-kit ⁺ Sca-1 ⁺	None	7,000–37,000	1–119	0/9
MHC-EGFP	Lin ⁻ c-kit ⁺	MI	100,000	14	0/10
β-Act-EGFP	Lin ⁻ c-kit ⁺	MI	50,000	7–14	0/27

MI, myocardial infarct.

ment into a surrogate cardiomyogenic developmental field. In other studies, co-cultures were performed using variable inputs of Lin⁻ c-kit⁺ cells from MHC-nLAC mice and embryonic day (E)15 fetal cardiomyocytes from non-transgenic mice. A similar approach was used to monitor transdifferentiation of adult endothelial progenitor cells²³. The cultures were fixed after 7 days and stained with X-gal. In five independently established co-cultures, no blue nuclei were observed.

Modification of the sorting protocol was used to assess the cardiomyogenic potential of other populations of bone-marrow-derived cells. Lin⁻ c-kit⁺ Sca-1⁺ stem cells (a population further enriched in multipotential primitive HSCs) prepared from MHC-nLAC transgenic mice were transplanted into normal ($n = 9$) or cautery injured ($n = 11$) congenic, non-transgenic recipients. No blue nuclei were observed after whole-mount X-gal staining of vibratome sections prepared from these hearts (Table 1). Additionally, Lin⁻ c-kit⁺ Sca-1⁺ cells prepared from MHC-nLAC transgenic mice were transplanted into normal ($n = 9$) or cautery injured

($n = 11$) congenic, non-transgenic recipients. Once again, no blue nuclei were detected in the recipient hearts (Table 1). Collectively these data suggest that HSCs do not undergo cardiomyogenic differentiation after transplantation into normal or injured hearts.

To rule out the possibility that the bacterial-derived β-galactosidase reporter gene might be subjected to excessive methylation when passed through non-cardiac cell lineages (that is, through the bone marrow), thus resulting in its silencing, additional experiments were performed using HSCs from mice with cardiac-restricted expression of the *Aequorea victoria* enhanced green fluorescent protein (MHC-EGFP mice). In isolated cell preparations, 100% of cardiac myocytes exhibit green fluorescence²⁴, and similar to the MHC-nLAC mice, the presence of even a single transplanted fetal cardiomyocyte was readily detected in control experiments. Using the same methodology as with MHC-nLAC donor cells, Lin⁻ c-kit⁺ HSCs from MHC-EGFP mice were transplanted at the peri-infarct zone of adult congenic, non-transgenic recipients. Of the ten recipient hearts assayed, none exhibited EGFP

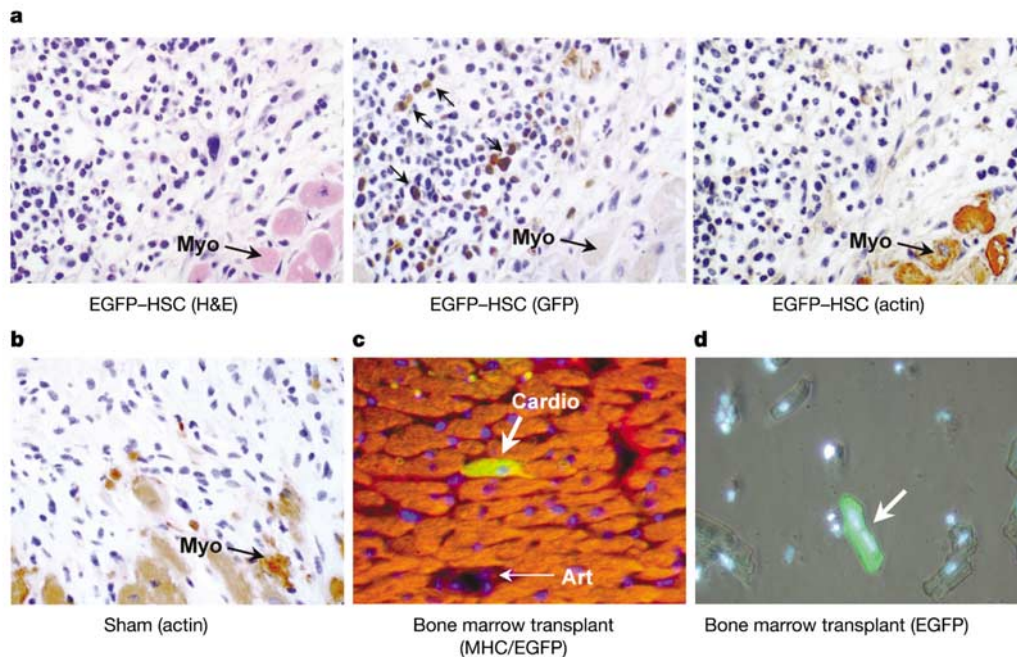


Figure 2 Absence of cardiac differentiation of HSCs after direct injection into infarcts, contrasted with rare transdifferentiation after bone marrow transplantation. β-Act-EGFP mice were cell donors. **a**, Left panel: haematoxylin- and eosin-stained section showing junction of host myocardium (Myo) with granulation tissue of 1-week-old, HSC-injected (EGFP-HSC) infarct. Granulation tissue contains numerous granulocytes and mononuclear inflammatory cells. Middle panel: serial section from the same heart immunostained for EGFP (brown), showing numerous EGFP⁺ cells dispersed throughout granulation tissue (arrows). Host myocardium (Myo) is unstained. Right panel: serial section from the same heart immunostained for sarcomeric actin (brown). Host myocardium (Myo) is strongly stained, but no sarcomeric actin is present in the region

containing EGFP-expressing cells. **b**, Sham-injected heart 1 week after infarction stained for sarcomeric actin (brown). Myocardium (Myo) at infarct border stains strongly, but infarct granulation tissue is negative. **c**, Histological detection of a rare EGFP⁺ cardiomyocyte (Cardio; yellow) in the peri-infarct region after bone marrow transplantation, shown by immunostaining for α-myosin heavy chain (red) and EGFP (green). Approximately 2–4 such cells were identified per heart. A small arteriole (Art) is unstained. **d**, Rare transdifferentiation event after bone marrow transplantation detected in enzymatically dispersed cardiomyocytes. A single rod-shaped cardiomyocyte contains EGFP (arrow), while multiple other cardiomyocytes are negative.

positivity (Table 1). Thus, two independent, cardiac-restricted reporter transgenes failed to demonstrate that Lin⁻c-kit⁺ stem cells undergo cardiomyogenic differentiation after transplantation into infarcted myocardium.

To follow more readily the fate of the transplanted Lin⁻c-kit⁺ stem cells, as well as to test a reporter transgene that uses a constitutively active promoter, additional experiments were performed using transgenic mice in which EGFP was ubiquitously expressed from the chicken β -actin promoter (β -Act-EGFP mice). Lin⁻c-kit⁺ HSCs derived from the β -Act-EGFP mice were transplanted into the peri-infarct zone of 15 adult congenic, non-transgenic recipients. EGFP-expressing cells, identified either by intrinsic fluorescence or by anti-GFP immunostaining, were abundant at 1 week after infarction (Fig. 2a, middle panel) but were much fewer in number at 2 weeks. The EGFP cells were small, round and did not co-localize with regions of sarcomeric actin staining in serial sections (Fig. 2a, right panel). To test whether an immune reaction to EGFP²⁵ might have prevented detection of transdifferentiation, these experiments were repeated using non-transgenic nude mice as recipients of transgenic Lin⁻c-kit⁺ cells ($n = 12$). Once again, the EGFP-expressing cells were small, round and did not co-localize with sarcomeric actin or myosin staining. Thus, even when a ubiquitously expressed transgene was used to follow cell lineage, no evidence for cardiomyogenic differentiation of HSCs was detected.

Independent of transgenic analysis, the extensive regeneration of the infarct region reported previously by Orlic and colleagues⁹ should be detectable in our studies by straightforward histological evaluation. Sections from mice receiving Lin⁻c-kit⁺ cells (with either the MHC-nLAC or β -Act-EGFP transgenes) were compared to sham-injected mice at 1–2 weeks after infarction. All sections revealed changes typical for 1–2-week-old infarcts (Figs 1d and 2a, right panel, b): there was a rim of surviving subendocardial myocardium, patchy amounts of surviving subepicardial myocardium and unresorbed necrotic myocardium that decreased in size from 1–2 weeks. The region surrounding the necrotic zone consisted of typical granulation tissue evolving towards immature scar tissue, and this granulation tissue had only background levels of myosin staining. The actin- or myosin-stained hearts with HSC transplants were indistinguishable from hearts with sham transplants, with no evidence of regenerating myocardium (Table 1). In addition to documenting a failure of the Lin⁻c-kit⁺ cells to undergo cardiomyogenic differentiation, these data also indicate that the presence of the transplanted HSCs does not result in the overt recruitment of endogenous cardiomyogenic stem cells.

Finally, the ability of circulating, bone-marrow-derived cells to give rise to cardiomyocytes after myocardial infarction was tested. Unfractionated bone marrow from β -Act-EGFP transgenic mice was transplanted into lethally irradiated, non-transgenic recipients. Mice were subjected to coronary artery ligation 2 months after engraftment, at which time their peripheral blood leukocytes consisted of >90% EGFP-expressing and thus donor-derived cells. Their hearts were studied from 1 week to 2 months after myocardial infarction, using double-label immunohistochemistry for EGFP and α -myosin heavy chain. Cells with anti-EGFP and anti-cardiac α -myosin heavy chain immune reactivity were detected in the peri-infarct zone (Fig. 2c), albeit only very infrequently (on average only 2–4 cells per heart were detected). As an independent method to ensure that the EGFP signal truly resided in cardiomyocytes, hearts were enzymatically dissociated and studied by fluorescence microscopy. Once again, fluorescent cardiomyocytes were observed at a similarly rare frequency in the dissociated cell preparations (1–3 cells per 100,000 cardiomyocytes; Fig. 2d).

The data presented here collectively suggest that direct injection of HSCs into the mouse heart does not result in *de novo* cardiomyogenic events or tissue regeneration. No lineage-restricted reporter gene activity was observed after injection of MHC-nLAC and MHC-EGFP HSCs into normal or injured hearts, indicating that

the cardiac α -myosin heavy chain promoter is not activated in the transplanted cells. This view is supported by the absence of co-localization of EGFP and myosin or actin by immunostaining. Indeed, not a single cardiomyogenic event was detected in the 145 HSC transplants that were analysed. Rather, the implanted cells remained morphologically consistent with haematopoietic cells and decreased in number from 1–2 weeks. Importantly, the presence of bone-marrow-derived cardiomyocytes after bone marrow transplantation and infarction (Fig. 2c, d) suggests that transdifferentiation and/or fusion events can be detected using the methods employed here. The very low level of cardiomyocyte repopulation after bone marrow transplantation in the current study is consistent with the work of Jackson *et al.*⁸, who estimated that 0.02% of cardiomyocytes in infarcted hearts arose from bone marrow sources. These data are also in agreement with most of the analyses of chimaerism in transplanted human hearts¹², including our own, where 0.04% of cardiomyocytes originated from extra-cardiac sources¹⁰.

These results contrast with the work of Orlic *et al.*, who reported extensive cardiac regeneration after direct injection of Lin⁻c-kit⁺ cells into infarcts⁹. The basis for this discrepancy is not clear, as considerable care was exercised to reproduce the methods for stem cell isolation and transplantation used by Orlic and colleagues (see the Supplementary Information for a more detailed comparison of the methods used). Nonetheless there still could be subtle differences in the protocols; moreover, differences in trace components in the stem cell preparation might contribute to the differential outcomes. An alternative and perhaps more likely explanation for the discrepant results lies in the different assays used to detect cardiomyogenic differentiation. The study by Orlic and colleagues⁹ relied exclusively on immune fluorescence staining to track cell fate and to monitor cell differentiation after HSC transplantation. This approach requires the establishment of signal thresholds, above which cells are designated as positive for a given marker (as, for example, GFP and myosin immune fluorescence). Establishing thresholds in tissues with high levels of nonspecific autofluorescence, as is typically encountered in the infarcted heart, is by nature a subjective process. For this reason, the bulk of the experiments in the current study used transgenic markers of both lineage and phenotype, which have low background and hence are intrinsically less subjective than immunostaining.

The absence of cardiomyocytes with activated reporter transgenes after intracardiac injection of transgenic Lin⁻c-kit⁺ stem cells suggests that cell fusion does not occur after this form of HSC delivery. Fusion of bone marrow cells with host cardiomyocytes in uninjured hearts has been reported recently¹⁴. Fusion between cardiomyocytes and donor cells was also observed after systemic delivery of genetically labelled adult heart-derived stem cells²⁶. It is possible that cell fusion might underlie the apparent transdifferentiation events attributed to circulating bone-marrow-derived stem cells^{8,10,12,13,15,27}, as well as those observed after intracardiac injection of adult marrow-derived progenitors²⁸. The absence of overt fusion events after intracardiac transplantation of HSCs in the current study may reflect differences in the mode of injury, the mode of delivery, and/or intrinsic properties of the stem cells used.

The data presented here did not address the potential beneficial effects of HSC injection on ventricular function after myocardial injury. Rather, the data indicate that Lin⁻c-kit⁺ progenitor cells isolated according to our methodologies fail to undergo overt cardiomyogenic differentiation when transplanted into normal or injured hearts. In this regard, it is possible that the functional benefits observed by Orlic *et al.*⁹ resulted from a beneficial impact on left ventricular remodelling and/or angiogenesis, rather than myocardial regeneration. Indeed, it is apparent that cell transplantation can result in improved cardiac function in the absence of donor cell participation in a functional syncytium with the host heart (reviewed in ref. 21). Finally, it is worth noting that several

clinical trials of bone marrow progenitor cells for cardiac repair have been initiated over the last 2 yr^{16,17}. The failure of HSCs to contribute significantly to formation of new cardiomyocytes in the present study may call into question the mechanistic underpinnings of such trials. □

Methods

Isolation of bone-marrow-derived HSCs

Tibias, femurs and iliac crests were collected from MHC-nLAC, MHC-EGFP or β -Act-EGFP mice, crushed in PBS containing 0.1% BSA and filtered through a 40- μ m nylon mesh to obtain crude bone marrow. Crude marrow was then fractionated on Histopaque (1.083 g ml⁻¹, Sigma) at 740g for 25 min to collect low-density marrow cells from the interface. Both mature and immature haematopoietic cells were depleted from low-density marrow cells by pre-incubation with lineage-specific rat antibodies to murine CD4, CD8, Gr-1, B220 and Mac-1 (Pharmingen) and subsequently labelling with anti-rat IgG microbeads followed by magnetic cell sorting (MACS, Miltenyi Biotec). Briefly, unlabelled progenitor cells were separated from magnetically labelled low-density marrow cells on a column, which was placed in the magnetic field of a MACS separator. The magnetically labelled cells were retained in the column. Cells (Lin-depleted) present in the flow-through were pelleted by centrifugation at 435g for 5 min and incubated with c-kit (conjugated with fluorescein isothiocyanate or phycoerythrin, Pharmingen) and/or Sca-1 (conjugated with phycoerythrin, Pharmingen) antibodies and sorted by FACS for Lin⁻ c-kit⁺ Sca-1⁺, Lin⁻ c-kit⁻ Sca-1⁺ or Lin⁻ c-kit⁺ cell types.

Coronary artery ligation and intracardiac grafting

This model was performed as detailed previously^{19,29}. Briefly, for studies involving stem cells from cardiac-restricted transgenic mice, recipient male and female mice were anaesthetized, supported on a ventilator, their left anterior descending coronary arteries ligated, and their chests closed aseptically. Five hours after ligation, mice were again anaesthetized, intubated, ventilated, and had their hearts exposed as above. The cell suspension was injected directly into the peri-infarcted area of the left ventricular free wall, as indicated in Table 1, using a 27 or 30 gauge needle. Sham-engrafted animals received comparable injections of serum-free medium. Closure and recovery were as above. All grafting experiments were done into histocompatible recipient mice such that no immune suppression was needed. Studies involving stem cells from β -Act-EGFP mice were performed as above, except that the HSCs were injected immediately after coronary ligation.

Histology

Histological methods are detailed in the Supplementary Information. For detection of LacZ reporter activity, hearts were fixed, vibratome sectioned at 300 μ m from apex to base, and whole-mount stained with X-gal substrate as described¹⁸. The sections were then carefully examined under a stereomicroscope for the presence of blue nuclei, a procedure capable of detecting a single positively stained nucleus in a heart²¹. The whole-mount sections were subsequently paraffin-embedded. Immunostaining for sarcomeric myosin heavy chain and sarcomeric actin were performed as previously described^{10,30}.

Chimaeric embryo bodies

HSCs (Lin⁻ c-kit⁺) were isolated from MHC-nLAC mouse bone marrow and mixed with undifferentiated mouse embryonic stem cells at 1:1, 1:2 and 1:8 ratios. Chimaeric embryo bodies were formed as detailed in the Supplementary Information. Embryo bodies were studied by X-gal staining and PCR analysis after 7–10 d of differentiation, when areas of spontaneous beating activity were present.

Bone marrow transplantation studies

Our bone marrow transplant protocol is detailed in the Supplementary Information. Wild-type C57Bl6/J mice were lethally irradiated and rescued by administration of ten million unfractionated bone marrow mononuclear cells obtained from β -Act-EGFP transgenic mice ($n = 13$). Myocardial infarction was performed 8–10 weeks post-transplant, when all animals showed >90% EGFP⁺ cells in peripheral blood. Mice were killed from 2–10 weeks after infarction and studied by immunostaining of tissue sections or microscopic analysis of enzymatically dispersed cells.

Received 20 October 2003; accepted 1 March 2004; doi:10.1038/nature02446.

Published online 21 March 2004.

- Mallory, G. K., White, P. D. & Salcedo-Salgar, J. The speed of healing of myocardial infarction: A study of the pathologic anatomy in 72 cases. *Am. Heart J.* **18**, 647–671 (1939).
- Ferrari, G. *et al.* Muscle regeneration by bone marrow-derived myogenic progenitors. *Science* **279**, 1528–1530 (1998); erratum *Science* **281**, 923 (1998).
- Gussoni, E. *et al.* Dystrophin expression in the mdx mouse restored by stem cell transplantation. *Nature* **401**, 390–394 (1999).
- Lagasse, E. *et al.* Purified hematopoietic stem cells can differentiate into hepatocytes *in vivo*. *Nature Med.* **6**, 1229–1234 (2000).
- Krause, D. S. *et al.* Multi-organ, multi-lineage engraftment by a single bone marrow-derived stem cell. *Cell* **105**, 369–377 (2001).
- Mezey, E., Chandross, K. J., Harta, G., Maki, R. A. & McKercher, S. R. Turning blood into brain: cells bearing neuronal antigens generated *in vivo* from bone marrow. *Science* **290**, 1779–1782 (2000).
- Brazelton, T. R., Rossi, F. M., Keshet, G. I. & Blau, H. M. From marrow to brain: expression of neuronal phenotypes in adult mice. *Science* **290**, 1775–1779 (2000).
- Jackson, K. A. *et al.* Regeneration of ischemic cardiac muscle and vascular endothelium by adult stem cells. *J. Clin. Invest.* **107**, 1395–1402 (2001).

- Orlic, D. *et al.* Bone marrow cells regenerate infarcted myocardium. *Nature* **410**, 701–705 (2001).
- Laflamme, M. A., Myerson, D., Saffitz, J. E. & Murry, C. E. Evidence for cardiomyocyte repopulation by extracardiac progenitors in transplanted human hearts. *Circ. Res.* **90**, 634–640 (2002).
- Quaini, F. *et al.* Chimerism of the transplanted heart. *N. Engl. J. Med.* **346**, 5–15 (2002).
- Muller, P. *et al.* Cardiomyocytes of noncardiac origin in myocardial biopsies of human transplanted hearts. *Circulation* **106**, 31–35 (2002).
- Bittner, R. E. *et al.* Recruitment of bone-marrow-derived cells by skeletal and cardiac muscle in adult dystrophic mdx mice. *Anat. Embryol. (Berl.)* **199**, 391–396 (1999).
- Alvarez-Dolado, M. *et al.* Fusion of bone-marrow-derived cells with Purkinje neurons, cardiomyocytes and hepatocytes. *Nature* **425**, 968–973 (2003).
- Deb, A. *et al.* Bone marrow-derived cardiomyocytes are present in adult human heart: A study of gender-mismatched bone marrow transplantation patients. *Circulation* **107**, 1247–1249 (2003).
- Strauer, B. E. *et al.* Repair of infarcted myocardium by autologous intracoronary mononuclear bone marrow cell transplantation in humans. *Circulation* **106**, 1913–1918 (2002).
- Assmus, B. *et al.* Transplantation of progenitor cells and regeneration enhancement in acute myocardial infarction (TOPCARE-AMI). *Circulation* **106**, 3009–3017 (2002).
- Soonpaa, M. H., Koh, G. Y., Klug, M. G. & Field, L. J. Formation of nascent intercalated disks between grafted fetal cardiomyocytes and host myocardium. *Science* **264**, 98–101 (1994).
- Virag, J. I. & Murry, C. E. Myofibroblast and endothelial cell proliferation during murine myocardial infarct repair. *Am. J. Pathol.* **163**, 2433–2440 (2003).
- Soonpaa, M. H. & Field, L. J. Assessment of cardiomyocyte DNA synthesis in normal and injured adult mouse hearts. *Am. J. Physiol.* **272**, H220–H226 (1997).
- Dowell, J. D., Rubart, M., Pasumarthi, K. B., Soonpaa, M. H. & Field, L. J. Myocyte and myogenic stem cell transplantation in the heart. *Cardiovasc. Res.* **58**, 336–350 (2003).
- Clarke, D. L. *et al.* Generalized potential of adult neural stem cells. *Science* **288**, 1660–1663 (2000).
- Badorff, C. *et al.* Transdifferentiation of blood-derived human adult endothelial progenitor cells into functionally active cardiomyocytes. *Circulation* **107**, 1024–1032 (2003).
- Rubart, M. *et al.* Physiological coupling of donor and host cardiomyocytes after cellular transplantation. *Circ. Res.* **92**, 1217–1224 (2003).
- Stripeck, R. *et al.* Immune response to green fluorescent protein: implications for gene therapy. *Gene Ther.* **6**, 1305–1312 (1999).
- Oh, H. *et al.* Cardiac progenitor cells from adult myocardium: homing, differentiation, and fusion after infarction. *Proc. Natl Acad. Sci. USA* **100**, 12313–12318 (2003).
- Kuramochi, Y. *et al.* Cardiomyocyte regeneration from circulating bone marrow cells in mice. *Pediatr. Res.* **54**, 319–325 (2003).
- Kudo, M. *et al.* Implantation of bone marrow stem cells reduces the infarction and fibrosis in ischemic mouse heart. *J. Mol. Cell. Cardiol.* **35**, 1113–1119 (2003).
- Reinecke, H. & Murry, C. E. Cell grafting for cardiac repair. *Methods Mol. Biol.* **219**, 97–112 (2003).
- Reinecke, H., Poppa, V. & Murry, C. E. Skeletal muscle stem cells do not transdifferentiate into cardiomyocytes after cardiac grafting. *J. Mol. Cell. Cardiol.* **34**, 241–249 (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements C.E.M. and L.J.F. thank L. Reinlib for his longstanding support of this collaboration. We thank C. Storey for assistance in sorting HSCs and in bone marrow transplantation, and L. Fernando Santana for assistance with enzymatic dissociation of mouse hearts. These studies were supported in part by NIH grants to C.E.M. and L.J.F., and by the HHMI (G.B., D.A.W.).

Competing interests statement The authors declare competing financial interests: details accompany the paper on www.nature.com/nature.

Correspondence and requests for materials should be addressed to C.E.M. (murry@u.washington.edu) or L.J.F. (ljfield@iupui.edu).

Haematopoietic stem cells adopt mature haematopoietic fates in ischaemic myocardium

Leora B. Balsam¹, Amy J. Wagers^{2,3}, Julie L. Christensen^{2,3}, Theo Kofidis¹, Irving L. Weissman^{2,3} & Robert C. Robbins¹

¹Departments of Cardiothoracic Surgery, ²Pathology, and ³Developmental Biology, Stanford University School of Medicine, Stanford, California 94305, USA

Under conditions of tissue injury, myocardial replication and regeneration have been reported¹. A growing number of investigators have implicated adult bone marrow (BM) in this process, suggesting that marrow serves as a reservoir for cardiac precursor cells^{2–6}. It remains unclear which BM cell(s) can contribute to myocardium, and whether they do so by transdifferentiation or cell fusion. Here, we studied the ability of c-kit-enriched BM cells,

Table 1 Estimated number of GFP⁺ cells present in injured myocardium

Animal number	Cell type	Original no. of GFP ⁺ cells implanted	Time of death (days)	Estimated number of GFP ⁺ cells at death	Original GFP ⁺ cells present at death (%)
1	c-kit ^{Enr}	1 × 10 ⁶	10	190,430	19.0
2	c-kit ^{Enr}	1 × 10 ⁶	10	471,930	47.2
3	c-kit ^{Enr}	1 × 10 ⁶	30	3,600	0.4
4	c-kit ^{Enr}	1 × 10 ⁶	30	3,000	0.3
5	c-kit ^{Enr}	1 × 10 ⁶	30	0	0
6	c-kit ^{Enr}	1 × 10 ⁶	30	0	0
7	KTLS LT-HSC	4 × 10 ³	10	25,980	649.5
8	KTLS LT-HSC	4 × 10 ³	10	11,820	295.5
9	KTLS LT-HSC	4 × 10 ³	10	16,440	411.0
10	KTLS LT-HSC	4 × 10 ³	30	0	0
11	KTLS LT-HSC	4 × 10 ³	30	0	0
12	KTLS LT-HSC	4 × 10 ³	30	720	18.0
13	Lin ⁻ c-kit ⁺	6 × 10 ⁵	10	61,800	10.3
14	Lin ⁻ c-kit ⁺	6 × 10 ⁵	10	64,890	10.8

Lin⁻ c-kit⁺ BM cells and c-kit⁺ Thy1.1^{lo} Lin⁻ Sca-1⁺ long-term reconstituting haematopoietic stem cells to regenerate myocardium in an infarct model. Cells were isolated from transgenic mice expressing green fluorescent protein (GFP) and injected directly into ischaemic myocardium of wild-type mice. Abundant GFP⁺ cells were detected in the myocardium after 10 days, but by 30 days, few cells were detectable. These GFP⁺ cells did not express cardiac tissue-specific markers, but rather, most of them expressed the haematopoietic marker CD45 and myeloid marker Gr-1. We also studied the role of circulating cells in the repair of

ischaemic myocardium using GFP⁺–GFP⁻ parabiotic mice. Again, we found no evidence of myocardial regeneration from blood-borne partner-derived cells. Our data suggest that even in the microenvironment of the injured heart, c-kit-enriched BM cells, Lin⁻ c-kit⁺ BM cells and c-kit⁺ Thy1.1^{lo} Lin⁻ Sca-1⁺ long-term reconstituting haematopoietic stem cells adopt only traditional haematopoietic fates.

A study published by Orlic *et al.*⁷ reported that an adult BM population enriched for haematopoietic stem cells (HSCs) (Lin⁻ c-kit⁺) transforms into myocardium and supporting vasculature within 9 days after direct injection into the ischaemic heart. To further characterize candidate cardiomyogenic populations within this BM subset, we isolated c-kit-enriched (c-kit^{Enr}), Lin⁻ c-kit⁺ and c-kit⁺ Thy1.1^{lo} Lin⁻ Sca-1⁺ (KTLS) long-term reconstituting HSCs (LT-HSCs) from GFP-transgenic donor mice. Wild-type recipient mice underwent ligation of the left anterior descending coronary artery, and shortly thereafter, donor BM cells were injected into the border zone of the ischaemic myocardium. Animals were killed at 10 or 30 days after infarction (Supplementary Fig. 1) to evaluate the presence of GFP⁺ donor-derived cells. At 10 days after infarction, large numbers of GFP⁺ cells were present in c-kit^{Enr} ($n = 6$), Lin⁻ c-kit⁺ ($n = 2$) and KTLS LT-HSC ($n = 4$) injected hearts. These cells were located primarily in clusters throughout the infarcted tissue and in the bordering peri-infarct zone. At 30 days after infarction, few (if any) GFP⁺ cells were detected in c-kit^{Enr} ($n = 5$) or KTLS LT-HSC ($n = 4$) injected hearts. The total number of GFP⁺ cells per heart was estimated in a subset of animals (Table 1). For recipients of c-kit^{Enr} cells, $33.1 \pm 19.9\%$ of the original number of injected GFP⁺ cells were detectable at 10 days, whereas only $0.2 \pm 0.2\%$ were noted at 30 days. Recipients of Lin⁻ c-kit⁺ cells contained $10.6 \pm 0.4\%$ of the original number of cells at 10 days. Recipients of KTLS LT-HSCs contained $452.0 \pm 181\%$ and $6.0 \pm 10.4\%$ of the original number of cells at 10 and 30 days, respectively. The data suggest that, at least in the case of injected KTLS LT-HSCs, cell division/proliferation may occur in the early time period after intramyocardial injection. However, over time, most of the cells in each group examined either die or migrate elsewhere.

We next performed two parallel experiments. In the first, donor cells (Lin⁻ c-kit⁺ or KTLS LT-HSCs) were injected into the myocardium 3–5 h after coronary ligation; in the second, donor cells (c-kit^{Enr} or KTLS LT-HSCs) were injected into the myocardium immediately after coronary ligation. We evaluated the phenotype of the intramyocardial GFP⁺ cells in each treatment group by standard and laser-scanning confocal microscopy at 10 days, and in some cases, 30 days after infarction. Cardiac tissue was stained with myocyte-specific, smooth-muscle-specific and endothelial-specific antibodies and sections were analysed for potential transdifferentiating GFP⁺ cells. As in our first study, there were many GFP⁺ cells, but none of these co-expressed cardiac tissue-specific markers (Fig. 1a–l). GFP⁺ cells also did not stain with antibodies against

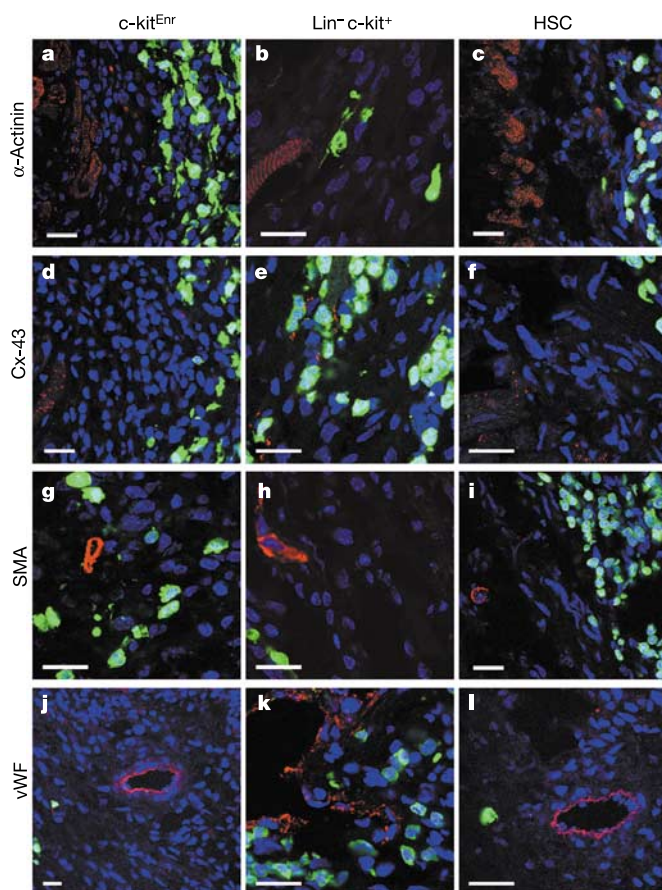


Figure 1 BM cells injected into ischaemic myocardium do not adopt cardiac, smooth muscle, or endothelial phenotypes. Representative confocal micrographs of infarct border zone 10 days after BM cell injection. **a–l**, DAPI nuclear staining is shown in blue, GFP expression in green, and tissue-specific markers in red (**a–c**, α -actinin; **d–f**, connexin-43; **g–i**, smooth muscle actin (SMA); **j–l**, von Willebrand's factor (vWF)). For panels **a, d, g, j** the donor BM cells were c-kit^{Enr}, for **b, e, h, k** they were Lin⁻ c-kit⁺, and for **c, f, i, l** they were purified KTLS LT-HSCs. Scale bar: 20 μ m.

c-kit, suggesting that c-kit⁺ LT-HSCs and haematopoietic progenitor cells within the originally injected cell populations had differentiated into alternate cell types. The pan-haematopoietic marker CD45 was co-expressed by many of the GFP⁺ cells in the c-kit^{Enr} treatment group at both 10 (Fig. 2a–d) and 30 days. At 10 days, the majority of these cells also expressed the granulocyte marker Gr-1 (Fig. 3a–d). In the Lin[−] c-kit⁺ and KTLS LT-HSC treatment groups, all GFP⁺ cells co-expressed CD45 (Fig. 2e–l) and Gr-1 at 10 days (Fig. 3e–l). At 30 days, rare GFP⁺ cells co-expressing the B-cell lymphoid marker B220 were also evident in the KTLS LT-HSC treatment group (Supplementary Fig. 2). No CD3-expressing (T-cell lineage) GFP⁺ cells were detected in the injected myocardium. Taken together, these results demonstrate that after injection into ischaemic myocardium, HSC-enriched Lin[−] c-kit⁺ cells and highly purified KTLS LT-HSCs differentiate exclusively into traditional haematopoietic cell fates, generating myeloid and lymphoid cell types. We did not observe a single case of transformation of these donor cells into cardiac myocytes, smooth muscle cells, or endothelial cells.

Although we found no evidence of myocardial transformation in cell-treated hearts, we hypothesized that cell treatment could still affect animal survival and preserve cardiac geometry and function after infarction. Peri-infarct treatment with c-kit^{Enr} BM cells ($n = 16$) or KTLS LT-HSCs ($n = 8$) did not improve 30-day survival after infarction when compared to control mice treated with saline only ($n = 12$). Survival was 87.5% in mice treated with c-kit^{Enr} cells, 87.5% in mice treated with KTLS LT-HSCs and 91.7% in control mice ($P =$ not significant). Infarct size and cardiac function were compared in a subset of mice treated with c-kit^{Enr} BM cells or saline only (control). Infarct size at 6 weeks after surgery was comparable between cell-treated and control mice (Supplemen-

tary Fig. 3). The percentage of left ventricle circumference below the left anterior descending artery ligature comprising scar tissue was $57.2 \pm 6.5\%$ in the cell-treated group ($n = 11$) and $59.6 \pm 9.6\%$ in the control group ($n = 7$) ($P =$ not significant). Parameters of cardiac geometry and function, including left ventricular chamber dimensions at end-diastole and end-systole, and per cent fractional shortening, were assessed *in vivo* by two-dimensional echocardiography at 2 and 6 weeks after infarction (Supplementary Table 1). At 2 weeks, a trend towards improved fractional shortening and decreased left ventricular chamber dimensions at end-diastole and end-systole was found in cell-treated versus control mice, yet these differences were not statistically significant; however, at 6 weeks, cell-treated mice showed a statistically significant, although modest, improvement in each of these parameters. Invasive haemodynamic parameters measured before death at 6 weeks after infarction did not differ significantly between cell-treated and control mice (Supplementary Table 2). Thus, we conclude that treatment with c-kit^{Enr} BM cells provides some long-term benefit in limiting ventricular dilatation and dysfunction after infarction, but it does not limit overall infarct size. A possible mechanism for this functional improvement could be potentiation of angiogenic activity of endogenous cells by injected c-kit^{Enr} cells.

Several studies have provided support for the hypothesis that BM cells are recruited to sites of cardiac injury through the blood and then transform into myocytes in the injured cardiac environment^{3,6}. In the non-ischaemic heart, we (A.J.W., J.L.C. and I.L.W.) found no KTLS LT-HSC-derived myocytes after haematopoietic reconstitution with a single GFP⁺ KTLS LT-HSC⁸. We also found no evidence of GFP⁺ cardiomyocytes in wild-type mice that had been joined by parabiosis to GFP-transgenic partners for 6–7 months. Nonetheless, we hypothesized that circulating cells, including BM-derived

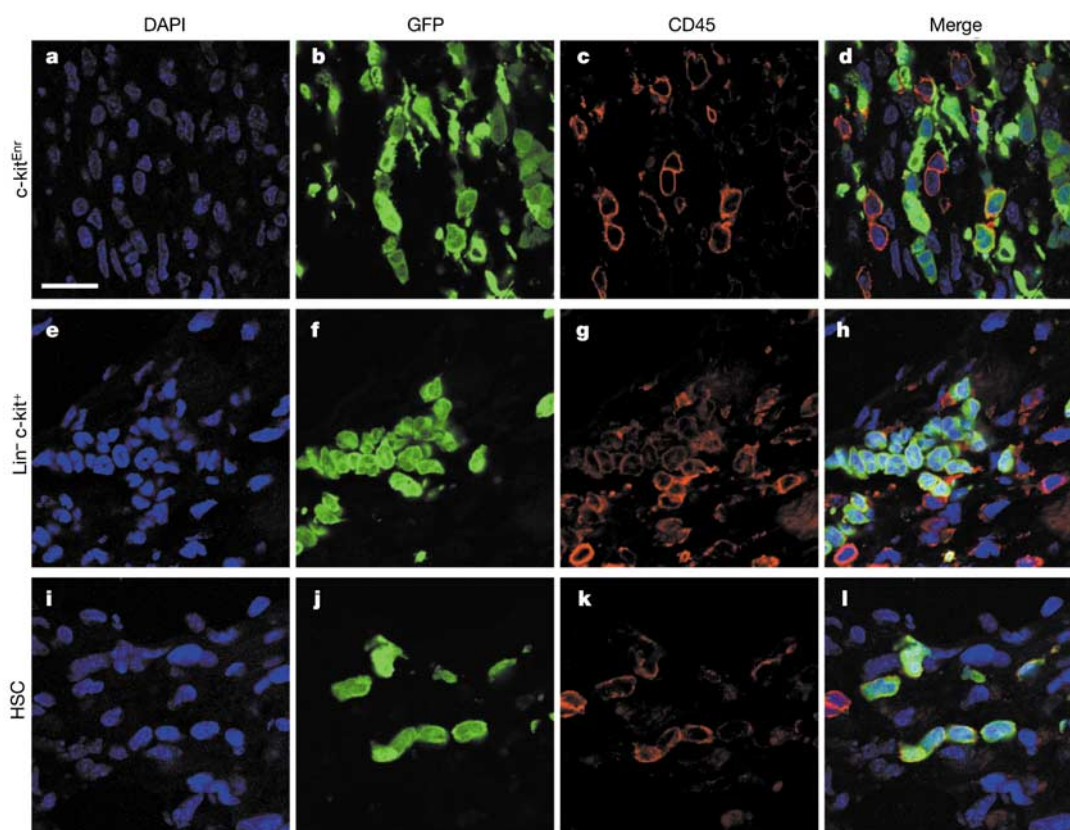


Figure 2 Haematopoietic phenotype of BM cells injected into ischaemic myocardium. **a–l**, Representative confocal micrographs of infarct 10 days after BM cell injection. DAPI nuclear staining (blue), GFP expression (green), CD45 expression (red), and merged

images are shown. For panels **a–d** the donor BM cells were c-kit^{Enr}; for **e–h** they were Lin[−] c-kit⁺; for **i–l** they were purified KTLS LT-HSCs. Scale bar: 20 μ m.

precursors, may transform into cardiomyocytes if significant tissue injury were present. To test this hypothesis, we used the classical parabiotic mouse model, in which surgically conjoined mice develop an anastomosed vasculature, leading to complete blood chimaerism within 7–10 days after surgery^{8,9}. Wild-type mice underwent left anterior descending artery ligation followed by immediate parabiosis to a GFP-transgenic partner. After 8 weeks, parabiotic pairs were killed and both haematopoietic (Supplementary Table 3) and cardiac chimaerism were assessed. Robust cross-engraftment of the blood, spleen and BM occurred in the wild-type partners, with GFP⁺ cells representing $43.5 \pm 4.6\%$, $38.1 \pm 3.2\%$ and $6.7 \pm 2.4\%$, respectively, of the total tissue cellularity. As described previously⁹, cross-engraftment of KTLS LT-HSCs was also evident, with $5.6 \pm 1.6\%$ GFP⁺ KTLS LT-HSCs present in the BM of non-transgenic parabiotic partners. Cardiac tissue of each wild-type parabiotic partner was assessed for the presence of GFP⁺ cells. Scattered GFP⁺ cells were present throughout the myocardium and scar; all of these cells expressed CD45 (Fig. 4a–h) and many co-expressed the B-cell marker B220 (Fig. 4i–p). Very rare cells co-expressed the T-cell marker CD3 (Fig. 4q–t). No co-expression of GFP with cardiac or smooth muscle markers was noted. These data suggest that even after cardiac injury, circulating cells do not detectably regenerate myocardium. Alternatively, it remains possible that regenerative cells circulate to the myocardium only early after myocardial injury, and such an event would be detected only if haematopoietic chimaerism were already established at the time of injury.

Recent years have seen tremendous excitement and controversy in the fields of stem cell biology and cardiac regeneration, with several studies suggesting that BM hosts precursors for cardiomyocytes. These studies were originally taken as evidence for trans-

differentiation, although more recent studies suggest that cell fusion, rather than transdifferentiation, probably underlies such observations^{10–13}. Here we demonstrate that three different populations of BM cells do not differentiate into cardiac muscle or vasculature when injected directly into ischaemic myocardium. Instead, they differentiate into mature haematopoietic lineages. Numerous donor BM-derived cells, many of them myeloid cells, are present in the heart at early time points, but by 1 month after injection, most of these cells are no longer detectable. Given the short half-life of myeloid cells, it is likely that these cells die in the intervening period of time.

Our studies are a logical extension to previous studies suggesting that BM cells contribute to myocardium after cardiac injury. Several studies have reported myocardial regeneration after transplantation in humans^{2,14}. Using the Y chromosome as a marker of recipient cells in male recipients of female donor hearts, studies report anywhere from 0.01% to 20% recipient-derived cardiomyocytes. Another study demonstrated cardiac chimaerism of 0.23% in the hearts of female patients that underwent BM transplantation from male donors⁵. Experimental models using BM chimaeric mice provide evidence for contributions of BM-derived cells to cardiomyocytes³ and suggest that direct fusion of blood lineage cells with myocytes is probably the mechanism by which such contributions occur¹². Here we extend these findings by injecting a variety of stem-cell-enriched populations directly into injured myocardium. Our results suggest that intramyocardial injection of these cells bypasses important events that occur during BM transplantation into lethally irradiated recipients that are critical for the contribution of these cells, by cell fusion or other mechanisms, to myocardial cell lineages.

It is difficult to reconcile our data with that of Orlic *et al.*⁷. There

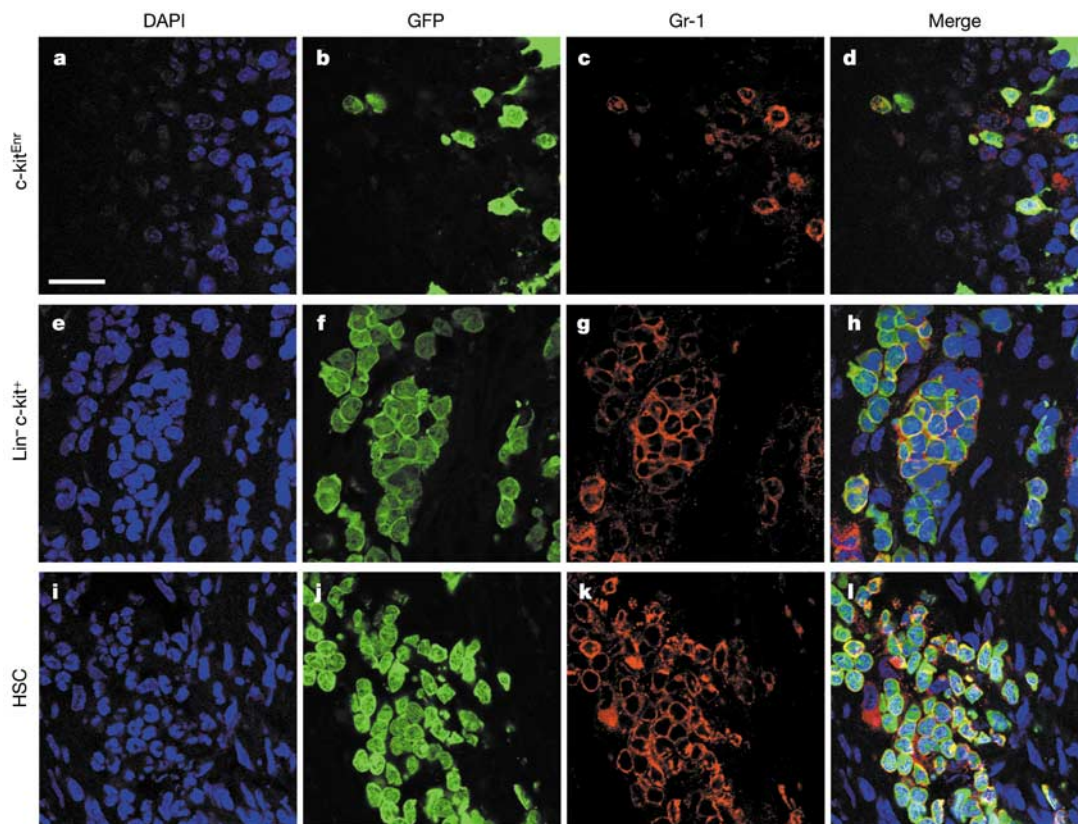


Figure 3 BM cells injected into ischaemic myocardium differentiate into granulocytes. **a–l**, Representative confocal micrographs of infarct 10 days after BM cell injection. DAPI nuclear staining (blue), GFP expression (green), myeloid Gr-1 expression (red), and

merged images are shown. For panels **a–d** the donor BM cells were c-kit^{Enr}; for **e–h** they were Lin[−] c-kit⁺; for **i–l** they were purified KTLS LT-HSCs. Scale bar: 20 μ m.

may be differences in our anaesthetic and/or surgical technique, and one may postulate that these resulted in a different outcome. Notably, we were able to detect GFP⁺ cells after 10 days in 14 out of 14 animals, whereas their group detected them in only 12 out of 30 animals, attributing this either to technical difficulty in implanting cells into the beating heart or to rejection of male donor cells implanted into female hearts. In our studies, we did not uniformly use gender-specific recipients or donors, with the exception of studies done with Lin⁻ c-kit⁺ cells. In those experiments, females served as recipients of GFP⁺ Lin⁻ c-kit⁺ male donor cells. Regardless,

in all 14 animals examined at the same 10-day post-infarct time point described by Orlic *et al.*⁷, we detected abundant GFP⁺ cells. This enhanced capacity for detection of GFP⁺ cells may reflect better technique during the injection step or more rigorous analysis of the tissue after harvesting. Consistent with Orlic *et al.*⁷, however, we do find that c-kit⁺ GFP⁺ input cells lose expression of the c-kit marker over time. The previous study noted only two c-kit⁺ cells in the harvested tissue. Unfortunately, they do not report staining for additional haematopoietic markers, such as CD45 or lineage-specific blood cell markers.

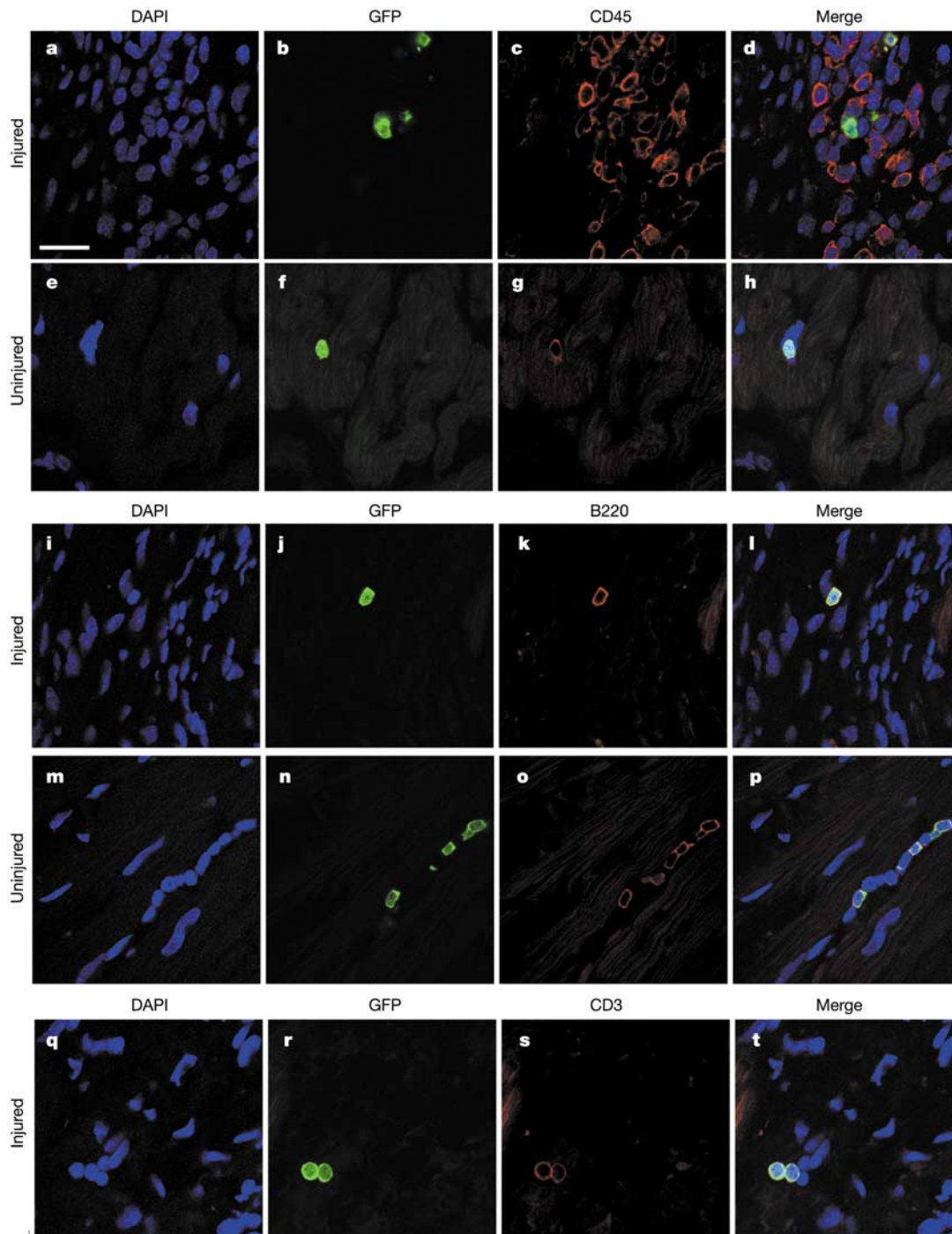


Figure 4 Chimaerism of cardiac tissue in non-transgenic parabiotic partners 8 weeks after cardiac injury and parabiosis. **a–t**, Representative confocal micrographs of left ventricular scar (injured area) and healthy left ventricular myocardium (uninjured area).

DAPI nuclear staining (blue), GFP expression (green), haematopoietic marker (CD45, B220, or CD3) expression (red), and merged images are shown. Scale bar: 20 μm.

Our data are of particular importance given the fact that a variety of groups worldwide have initiated clinical trials of BM transplantation into ischaemic myocardium^{15,16}. Without additional pre-clinical experimental data, these studies are premature and may in fact place a group of sick patients at risk. At the very least, many more preclinical experimental data should be collected before such phenomena can be fully understood or clinically exploited. □

Methods

Animals

Eight- to twelve-week-old C57BL/Ka-Thy-1.1 and enhanced GFP-transgenic C57BL/Ka-Thy-1.1 mice were bred and maintained at the Stanford University Research Animal Facility. Transgenic mice were generated as previously described, with GFP expression driven constitutively by the chicken β -actin promoter¹⁷. All animal procedures were approved by an institutional review board.

Cell purification

BM cells were isolated from GFP-transgenic mice and enriched for c-kit by magnetic cell sorting as described⁸. KTLS LT-HSC or Lin[−] c-kit⁺ populations were further purified by double fluorescence-activated cell sorting. The antibodies used included 19XE5 (anti-Thy1.1, phycoerythrin or fluorescein isothiocyanate conjugate), 2B8 (anti-c-kit, APC conjugate) and E13-161.7 (anti-Sca-1, Ly6A/E, Texas red conjugate). The cocktail of lineage marker antibodies included KT31.1 (anti-CD3), GK1.5 (anti-CD4), 53-7.3 (anti-CD5), 53-6.7 (anti-CD8), Ter119 (anti-erythrocyte-specific antigen), 6B2 (anti-B220), 8C5 (anti-Gr-1) and M1/70 (anti-Mac-1). These antibodies were produced and purified in the I.L.W. laboratory.

Determination of haematopoietic chimaerism

Parabiotic partners were screened by flow cytometry for blood, spleen and BM chimaerism 8 weeks after surgery as described previously⁸.

Myocardial injury model

C57BL/Ka-Thy-1.1 mice were endotracheally intubated and ventilated with a rodent ventilator (Harvard Apparatus). Anaesthesia was maintained with inhalational isoflurane. A thoracotomy was performed and an 8-0 ethilon suture was placed around the left anterior descending artery just below the atrioventricular border. BM cells (1×10^6 c-kit^{Enr}, 6×10^5 Lin[−] c-kit⁺, or 4×10^3 KTLS LT-HSCs) or saline were injected into the anterior and posterior infarct border zones. The chest was then closed and the animals were weaned from the ventilator and extubated. Animals were killed at 10 days, 30 days, or 6 weeks after injury.

Parabiosis and myocardial injury model

In these experiments, C57BL/Ka-Thy-1.1 mice underwent left anterior descending artery ligation followed immediately by surgical parabiosis to a GFP-transgenic C57BL/Ka-Thy-1.1 partner. Parabiosis was performed as described previously^{8,9}. Parabiotic pairs were housed individually with free access to food and water and were killed after 8 weeks.

Tissue collection and immunofluorescent histology

Hearts were perfused with saline and rapidly excised. They were fixed in 2% paraformaldehyde and cryoprotected in 30% sucrose. Tissue was frozen in optimum cutting temperature compound and stored at -70°C until sectioning. Immunostaining was performed on 5- μm sections with a panel of cardiac, smooth muscle, endothelial and haematopoietic antibodies. Primary antibodies included rabbit anti-connexin-43 antibody, mouse monoclonal anti- α -actinin antibody, mouse monoclonal anti-smooth muscle actin antibody (all Sigma), sheep anti-von Willebrand's factor antibody (Serotec), rat anti-c-kit antibody (Research Diagnostics), rat anti-Gr-1 antibody, rat anti-CD45 antibody (both BD Pharmingen), rat anti-CD3 antibody, rat anti-B220 antibody, goat anti-GFP antibody (Rockland), and rabbit anti-GFP Alexa Fluor 488 conjugated antibody (Molecular Probes). Texas red conjugated secondary antibodies were used for cardiac (connexin-43, α -actinin), smooth muscle (smooth muscle actin), endothelial (vWF) and haematopoietic primaries (c-kit, Gr-1, CD45, CD3, B220), and fluorescein isothiocyanate conjugated secondary antibody was used for the goat anti-GFP primary. Sections were counterstained with 4,6-diamidino-2-phenylindole (DAPI) and analysed with a Zeiss LSM 510 two-photon confocal laser-scanning microscope.

Estimate of GFP⁺ cells in post-infarcted hearts

The number of GFP⁺ cells in the myocardium was estimated at 10 and 30 days after surgery in a subset of animals. Fixed hearts were sectioned at a thickness of 5 μm from apex to base on a cryostat. Every 150 μm (that is, every 30th tissue section), the number of GFP⁺ cells was counted. Assuming a cell thickness of 5 μm , the total number of GFP⁺ cells per heart was estimated as 30 times the total number of GFP cells on all sections sampled.

Calculation of infarct size

Infarct size was calculated in a subset of animals 6 weeks after surgery using computer-

based planimetry. Fixed hearts were sectioned into six thick slices from base to apex, beginning at the level of the ligature on the left anterior descending artery. Thin sections from each level were stained with Masson's trichrome stain (Sigma). Infarct size was defined as the sum of the epicardial and endocardial infarct circumference divided by the sum of the total left ventricular epicardial and endocardial circumferences, and was averaged across the six slices.

Echocardiography

Two-dimensional echocardiography was performed at 2 and 6 weeks after surgery in cell-treated (1×10^6 c-kit^{Enr} cells) infarcted animals and saline-treated infarcted animals. Mice were anaesthetized with 2% inhalational isoflurane and imaged with a 15.8-MHz probe connected to a Sequoia C256 echocardiographic machine (Acuson).

Invasive haemodynamics

Closed-chest invasive haemodynamics were measured at 6 weeks after surgery in cell-treated (1×10^6 c-kit^{Enr} cells) infarcted ($n = 8$) and saline-treated infarcted animals ($n = 6$) with a Millar SPR-671 Mikro-tip pressure transducer (Millar Instruments). Mice were anaesthetized with 2% inhalational isoflurane. Recordings were analysed with Sonosoft 3.2.2 software (Sonometrics).

Statistics

Results are mean $\pm$ standard deviation. Survival was compared with the log-rank test. Comparisons for functional studies were made using the two-tailed Student's *t*-test. $P < 0.05$ was considered significant.

Received 21 September 2003; accepted 3 March 2004; doi:10.1038/nature02460.

Published online 21 March 2004.

- Beltrami, A. P. *et al.* Evidence that human cardiac myocytes divide after myocardial infarction. *N. Engl. J. Med.* **344**, 1750–1757 (2001).
- Quaini, F. *et al.* Chimerism of the transplanted heart. *N. Engl. J. Med.* **346**, 5–15 (2002).
- Jackson, K. A. *et al.* Regeneration of ischemic cardiac muscle and vascular endothelium by adult stem cells. *J. Clin. Invest.* **107**, 1395–1402 (2001).
- Toma, C., Pittenger, M. F., Cahill, K. S., Byrne, B. J. & Kessler, P. D. Human mesenchymal stem cells differentiate to a cardiomyocyte phenotype in the adult murine heart. *Circulation* **105**, 93–98 (2002).
- Deb, A. *et al.* Bone marrow-derived cardiomyocytes are present in adult human heart: a study of gender-mismatched bone marrow transplantation patients. *Circulation* **107**, 1247–1249 (2003).
- Orlic, D. *et al.* Mobilized bone marrow cells repair the infarcted heart, improving function and survival. *Proc. Natl Acad. Sci. USA* **98**, 10344–10349 (2001).
- Orlic, D. *et al.* Bone marrow cells regenerate infarcted myocardium. *Nature* **410**, 701–705 (2001).
- Wagers, A. J., Sherwood, R. I., Christensen, J. L. & Weissman, I. L. Little evidence for developmental plasticity of adult hematopoietic stem cells. *Science* **297**, 2256–2259 (2002).
- Wright, D. E., Wagers, A. J., Gulati, A. P., Johnson, F. L. & Weissman, I. L. Physiological migration of hematopoietic stem and progenitor cells. *Science* **294**, 1933–1936 (2001).
- Terada, N. *et al.* Bone marrow cells adopt the phenotype of other cells by spontaneous cell fusion. *Nature* **416**, 542–545 (2002).
- Ying, Q. L., Nichols, J., Evans, E. P. & Smith, A. G. Changing potency by spontaneous fusion. *Nature* **416**, 545–548 (2002).
- Alvarez-Dolado, M. *et al.* Fusion of bone-marrow-derived cells with Purkinje neurons, cardiomyocytes and hepatocytes. *Nature* **425**, 968–973 (2003).
- Oh, H. *et al.* Cardiac progenitor cells from adult myocardium: homing, differentiation, and fusion after infarction. *Proc. Natl Acad. Sci. USA* **100**, 12313–12318 (2003).
- Laflamme, M. A., Myerson, D., Saffitz, J. E. & Murry, C. E. Evidence for cardiomyocyte repopulation by extracardiac progenitors in transplanted human hearts. *Circ. Res.* **90**, 634–640 (2002).
- Tse, H. F. *et al.* Angiogenesis in ischaemic myocardium by intramyocardial autologous bone marrow mononuclear cell implantation. *Lancet* **361**, 47–49 (2003).
- Stamm, C. *et al.* Autologous bone-marrow stem-cell transplantation for myocardial regeneration. *Lancet* **361**, 45–46 (2003).
- Wright, D. E. *et al.* Cyclophosphamide/granulocyte colony-stimulating factor causes selective mobilization of bone marrow hematopoietic stem cells into the blood after M phase of the cell cycle. *Blood* **97**, 2278–2285 (2001).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank G. Hoyt for technical assistance with animal surgery, and V. Mariano and L. Hildalgo for animal care. This work was supported by the Falk Cardiovascular research fund (R.C.R.) and an NIH grant (I.L.W.). L.B.B. was supported by Thoracic Surgery Foundation Nina Starr Braunwald Research Training Fellowship; A.J.W. was supported by an American Cancer Society grant and the Frederick Frank/Lehman Brothers, Inc. Irvington Institute Fellowship; J.L.C. was supported by a NIH Training Grant in Molecular and Cellular Immunobiology; and T.K. was supported by a German Research Society Training Grant.

Competing interests statement The authors declare competing financial interests: details accompany the paper on www.nature.com/nature.

Correspondence and requests for materials should be addressed to R.C.R. (robbins@stanford.edu).

Contacts

Publisher: Ben Crowe

Editor: Paul Smaglik

Marketing Manager: David Bowen

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Naturejobs Sales Director:

Nevin Bayoumi (4978)

UK/ RoW/ Ireland:

Matt Powell (4953)

Andy Douglas (4975)

Frank Phelan (4944)

Scandinavia/ Spain/ Portugal:

Evelina Rubio Håkansson (4973)

Natureevents: Silke Opstrup (4994)

France/ Switzerland:

Amelie Pequignot (4974)

Production Manager:

Billie Franklin
To send materials use London
address above.

Tel +44 (0) 20 7843 4814

Fax +44 (0) 20 7843 4996

e-mail: naturejobs@nature.com

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Stefan Hales

European Satellite Office

Germany/ Austria/ Italy/

The Netherlands/ Belgium:

Patrick Phelan, Odo Wulffen

Tel + 49 89 54 90 57 11/-2

Fax + 49 89 54 90 57 20

e-mail: p.phelan@nature.com

o.wulffen@nature.com

US Head Office, New York

345 Park Avenue South,
10th Floor, New York, NY 10010-1707
Tel +1 800 989 7718

Fax +1 800 989 7103

e-mail: naturejobs@natureny.com

US Sales Manager:

Peter Bless

Japan Head Office, Tokyo

MG Ichigaya Building (5F),

19-1 Haraikatamachi,

Shinjuku-ku,

Tokyo 162-0841

Tel +81 3 3267 8751

Fax +81 3 3267 8746

Asia-Pacific Sales Director:

Rinoko Asami

e-mail: rasami@naturejp.com

naturejobs

How to succeed in business

Decoding acronyms such as VC and IPO was once primarily the domain of a professional bearing another acronym — the MBA. Then, in the biotech boom that anticipated the sequencing of the human genome, life-science PhDs started schooling themselves. Now, even undergraduates are getting in on the act, taking courses tailored to teach a scientist how to think like a board chairman.

The evolution in business education has mirrored the involvement of scientists in business-plan competitions around the world (see page 676). But for scientists to excel in both science and business in a world where, like it or not, academic and industrial science are increasingly linked, young scientists could use more training to compete with their business-school colleagues.

California, with its globally dominant life-science hubs in San Francisco and San Diego and a growing medical-devices hub near Los Angeles, is one of the leaders in business training for scientists. The University of California at San Diego has for the past few years offered biotech business courses which are attended by biotech workers, graduate students and even visitors flying in from other continents. The University of San Francisco offers a course called 'From Idea to IPO' that walks life-science graduates through the whole business-plan process. And this spring, the University of California at Irvine launched an undergraduate entrepreneurship course for scientists and engineers. This filled two 40-student sessions to capacity, with a waiting list to boot.

The demand for such courses — whatever the training level — surely isn't limited to California, or to the life sciences. More offerings like these all over the world would help academic scientists speak the language of business. This is always a useful skill, even for those who plan to stay in the lab rather than move into the boardroom.

Paul Smaglik
Naturejobs editor



Contents

SPECIAL REPORT

Glittering prizes p676

CAREER VIEW

Scientists & Societies

Membership drive

Graduate Journal

Networking savvy

Movers

Zhenan Bao

p678

WWW.NATUREJOBS.COM

Career centre

Information on the
scientific job market

FOCUS

SPOTLIGHT

RECRUITMENT

ANNOUNCEMENTS

EVENTS

Entrepreneur contests spark innovation and provide a practice run for start-ups. Kendall Powell sizes up the competitions.

Sitting on a London bus, Donna Winderbank-Scott and two other medical students from Imperial College London dreamt up a device that may one day help millions of asthma patients. The motivation? A £1,000 (US\$1,800) prize awarded to the best one-page business idea submitted to Imperial's Ideas Challenge.

With colleagues Vanh Dang and Rachel Hames, Winderbank-Scott went on to win Imperial's New Business Challenge (£25,000) and the Idea 2 Product international business-plan competition at the University of Texas at Austin (US\$10,000), which also earned them a coveted spot in MOOT CORP, the 'World Cup' of business-plan competitions, also at the University of Texas.

Winderbank-Scott explains the brainstorm that generated the winning proposal: "What might have a big target population? Asthma. What are the problems in asthma treatment? Inhalers are difficult to use correctly. So we decided to try to make a better inhaler." Their product, which they named Inhalit, snaps on to any inhaler, improves drug delivery, monitors the severity of the disease, and alerts the patient if they need medical attention. The women expect to sell their idea to a company for royalties later this year.

Seems too simple? Well, there were "many, many meetings in various cafés" and "hundreds of e-mails" that Winderbank-Scott says went into transforming their idea into a full-blown business plan. "We had to learn everything from scratch: financing, marketing and advertising." Scientists who have participated in similar contests agree the most valuable part of the experience was gaining business savvy.

Business schools across the globe have been sponsoring competitions for their MBA students for some 15 years. The contests give the entrepreneur teams a practice trial by fire to pitch proposals to venture capitalists (VCs). They also encourage and steer the best innovations towards start-up companies.

During the past few years, competitions have begun to tap a new source of ideas — patents in university technology-transfer offices, often gathering dust. At the same time, more PhD students and researchers have begun participating, to develop their business ideas and pick up key contacts in the venture-funding world.

GETTING GRILLED

Several dozen universities around the world also hold international competitions, some of whose winners go on to take part in MOOT CORP — inspired by the 'moot courts' held at law schools. Top prizes range from \$2,000 (enough to get incorporated) to \$500,000.



DONNA WINDERBANK-SCOTT

In competition: (clockwise from above) Donna Winderbank-Scott (left) and friends developed an improved inhaler for asthma; the Watermark team from Bond University, Australia, present a new water treatment; and the ZORK team from Adelaide University have invented a better bottle-top.

Each team has 10–20 minutes to pitch, describing their product or technology, the market share, competitors in the area and how the product will be marketed. The presentation might conclude with financial projections and an exit strategy, or how the venture-capital firm will regain its investment.

Presentations are judged by a group of VCs, lawyers and accountants that would rival any thesis committee. They vote with their mock wallets. But young entrepreneurs say that the grilling session provides the most valuable feedback.

"The questions they raise are the fundamental ones you are going to have to answer later," says Sandra Waugh Ruggles, co-founder of Catalyst Biosciences in South San Francisco. She started her company with fellow PhDs from the University of California, San Francisco, after participating in several competitions.

A few competitions also give prizes for phone or 'elevator pitches' — condensed versions meant to catch the ear of a venture capitalist in two minutes. Sanford Ehrlich, who oversees the Venture Challenge at San Diego State University, says the Golden Phone Award was instituted on the advice of an investor who said he would not meet entrepreneurs who were unable to express ideas quickly on the phone.

Competitions offer practice for that real-life encounter, convincing a venture capitalist to invest precious dollars in your idea. Teams can submit a business plan based on original ideas for an on-paper-only company, or they can polish up the plan of a start-up company. Tom Kelly, founder and chief technical officer of Imago Scientific Instruments in Madison,



Wisconsin, has advised two winning teams on the business plan for his atomic imaging start-up company.

Kelly, a materials scientist, says his company benefited from the experience. The teams took a fresh look at the company's plan and found ways to increase their estimation of market share. But the real test was whether the judges at MOOT CORP would believe it.

"One VC knew our industry very well and said if our product got accepted as a tool, we had the potential to reach a billion-dollar market," Kelly recalls. "Prior to that, we hadn't heard the 'b-word'." The team came in second, and Kelly says the visibility helped boost fund-raising efforts. "We met with

several major VC groups as a result of that exposure."

The situation turned out to be win-win all around. Imago, which makes an atomic probe imaging instrument, raised \$12 million in two rounds of venture financing. The benefits of meeting VCs or successful entrepreneurs face-to-face cannot be overstated.

FROM LAB RAT TO CHIEF EXEC

While the competitions themselves are helpful, courses leading up to or spinning off from them also prove valuable, and are growing in popularity. Ruggles credits a competition lecture series and a course offered through the UCSF Center for Bioentrepreneurship for exposing her to the basics of business and finance. The UCSF course, called Idea to IPO, meets the needs of an increasing number of students who aim to enter the biotechnology industry.

Charles Craik, a chemical biologist who co-teaches the course with a Bay area venture capitalist, considers that preparing students for the business world is an important aspect of graduate education. The course grew in four years from 15 to 53 participants.

Workshops offered by many competitions may be even more valuable than the 15-minute final presentation. Singapore Management University (SMU) hosts the six finalist teams in the Lee Kuan Yew Global Business Plan Competition for three days of 'hothouse' sessions with business experts before the final showdown. And the Venture Bowl, hosted by the US National Institute for Entrepreneurship, awards the winning team with a week-long business 'boot camp' in New York City customized to launch their idea. Uniquely among the competitions, Venture Bowl also sets winners up with a potential contract for up to \$500,000 with the sponsoring VC firm, Carrot Capital.

HOME-GROWN ECONOMICS

Several countries are betting on university technology start-ups to stimulate growth and give a global edge. Grace Cheng, who coordinates the SMU competition, says it was created by the government to encourage young Singaporeans to start up the companies that form a cornerstone of the country's economy.

In Portugal, the commerce department and the biotechnology association, APBio, realized that many highly trained Portuguese scientists had too few options for industry jobs at home. So they started the Concurso de Ideias Bioempreendedor competition to encourage young bioentrepreneurs to start their own companies. Luis Amado, executive director of APBio and chief operating officer at Biotecnol in Oeiras, explains that the contest gives people tools "to solve this problem by themselves". The contest has helped at least four start-ups to put down roots in Portugal.

Contestants value the advantage that contests bring to developing a solid business proposal and learning the ropes in a no-cost dry run. And the potential to give voice to truly innovative ideas during contests looms large. Winderbank-Scott notes, "If you wanted to, you could come up with absolutely anything. It's really just thinking about a problem and coming up with a solution." Just like working on a thesis.

Kendall Powell is a freelance science writer based in Broomfield, Colorado.

Web links

MOOT CORP at the University of Texas, Austin
www.mootcorp.org/index.asp
 McGinnis Venture Competition, Carnegie Mellon University
www.mcginisventurecompetition.com/home.htm
 Venture Challenge, San Diego State University
www-rohan.sdsu.edu/dept/emc/venture_challenge
 UCSF's Center for BioEntrepreneurship
www.ucsf.edu/cbe
 Concurso Bioempreendedor in Portugal
www.bioempreendedor.com
 Imperial College London Entrepreneurs' Challenge
www.ec.ms.ic.ac.uk/challenge
 LKY Global Business Plan Competition at Singapore Management University
www.smu.edu.sg/lky
 Venture Bowl, National Institute for Entrepreneurship
www.venturebowl.com

GRADUATE JOURNAL

On the cards

Before I obtained my first set of business cards, I considered them to be accessories only for business people or perhaps the overly career savvy — certainly not necessary for a graduate student like myself. But I changed my mind when I became involved in the student association during my studies. I realized then that cards were an essential way to market the organization.

Now I see business cards as an essential tool in networking. I try to enhance their value by adding information; writing down the subject of the conversation on the card that accompanied its exchange is a good way to remember key points, and ensure you follow up with the person who gave you the card. And after follow-up, a marked-up card serves as a good archive of the initial meeting and actions taken as a result of it.

At conferences and workshops, it seems as though graduate students vary in their approach to business cards. Some have them and exchange them, others don't. I would encourage students to get a set printed — with a simple design and basic contact information. And finally, get comfortable approaching colleagues and swapping cards. ■

Philipp Angerer is a second-year PhD student in biotechnology at the Swiss Federal Institute of Technology (ETH) in Zurich, Switzerland.

SCIENTISTS & SOCIETIES

First, find your postdocs...

When I arrived at Georgetown University (GU) in Washington in 2001, I went to have my photo taken for a university discount card. The clerks asked my name, and whether I was a student. I told them I was a postdoctoral fellow and they responded with a blank stare. Then they asked me if I taught classes, which would qualify me as a faculty member. Clearly, this was the first that they had heard the word postdoc. This incident was telling. In starting up a postdoc organization, the most difficult task is simply finding the postdocs. Most are not considered to be 'real' university employees, so they are lost in translation somewhere between the accounting office and the lab, never making it to human resources.

After attending the inaugural National Postdoctoral Association (NPA) meeting in Berkeley, California, last year, I began establishing the Georgetown University Postdoctoral Association (GUPDA) with the help of my co-chair, Paul Lea. By far the most difficult task continues to be identifying the names of GU postdocs. To find them, we had to come up with some creative strategies.

We started with a list of 20 or so postdocs in my department and asked the university for enough money to hold a barbecue. We posted flyers that didn't give the location, but which asked for an RSVP. This way, we accumulated the e-mail addresses of a lot more postdocs. About 50 postdocs came, and we collected more e-mails.

One by one I added the names and e-mail addresses to an e-mail list server. Within three months the group had grown to more

than 200 postdocs, with half of them turning out for our biggest barbecue yet. Through networking, creating the e-mail list and continuing to post flyers for events, we have expanded our base.

Next we created a website, rallied support from faculty members and high-level administrators, and became a chartered organization. Now that we are 'official', we've found that the difficulty of finding all postdocs on campus rolls over to two other challenges — getting postdocs involved in the organization, and keeping it alive after Paul and I leave. Both could be even more difficult than knowing the exact number of postdocs at GU — a question that remains unanswered. ■

Lille Tidwell is a neuroscience postdoc at Georgetown University Medical Center and chair of the Georgetown University Postdoctoral Association.

♦ <http://gupda.georgetown.edu>

MOVERS Zhenan Bao, Associate Professor of Chemical Engineering, Stanford University, California



Arriving in the United States from China in 1990, Zhenan Bao felt a sense of urgency to complete her education. "As an immigrant, I wanted a US degree as soon as possible so I could start my career here." Since she had family in the Chicago area, she moved there. She already had a strong background in engineering from her education in China and an interest in polymers fostered by a summer programme at Nanjing University.

She learned that, with her existing education, she could get into graduate school at Chicago University without completing a BS. By a stroke of luck,

Luping Yu, an expert in polymer chemistry, had also recently arrived. Bao says she was fortunate because Yu was the only polymer chemist in the university at the time.

After her PhD with Yu, she had to choose between a temporary fellowship at the University of California, Berkeley, and a permanent position at Bell Labs. Bao's long-term goal was academia, but she decided on Bell Labs, with its basic research emphasis and history of publishing. "People told me, if you go to Bell labs, you can still go into academia later," she says.

Good mentoring again made the difference. At Bell Labs, manager Elsa Reichmanis, currently director of the materials research department, told her to take her time and "just talk to people and decide what you want to work on". That flexibility, friendly and helpful colleagues, and having a female mentor in a male-dominated discipline helped

Bao find her bearings. "It is important to see a woman being successful." Bao was impressed that Reichmanis was able to excel at Bell, gain membership of the US National Academy of Engineering, serve as the American Chemical Society president, and still raise four children. Other mentors at Bell Labs included Andy Lovinger, now at the polymer division of the National Science Foundation, and Ed Chandross.

Bao took Reichmanis's advice and soon found a project in plastic transistors where she could use her knowledge in a different application. "I thought my background would allow me to contribute to that project immediately," Bao says.

She didn't feel ready to move back to academia until a year ago. Now she is comfortable having made the jump and is guiding graduate students. And she aims to help them the way her many mentors helped her. ■

CV **1995–2003:** Research scientist, Bell Labs, Lucent Technologies, New Jersey
2001–present: Honorary professor, Department of Chemistry, East China University of Science and Technology, Shanghai
1995: University of Chicago, PhD, chemistry